cessda saw

Research and Innovation Action

CESSDA Strengthening and Widening

Project Number: 674939 Start Date of Project: 01/08/2015 Duration: 27 months

Deliverable D3.5 Report on the state-of-the-art, obstacles, models and roadmaps for widening the data perimeter of the data services

Dissemination Level	PU
Due Date of Deliverable	30/06/2017
Actual Submission Date	30/10/2017
Work Package	WP3 - Strengthening and widening through planning and engagement
Task	T3.4
Type	Report
EC Approval Status	Not approved yet
Version	V1.1
Number of Pages	p. 1 - p.137

Abstract: The task reviewed the state of play regarding specific data domains (data provided by academia, official statistics including administrative data, historical, health data and big data that means existing and emerging data types. Experiences and best practices are presented in this report with the objective of providing a practical roadmap, given that widening of CESSDA needs to address new data sources and new actors.

Project funded by the EU Horizon 2020 Research and Innovation Programme under the agreement No.674939

History

Version	Date	Reason	Revised by
0.1	20/12/2016	1st draft version by all partners	Nathalie Paton (CNRS)
0.2	30/03/2017	2nd draft version by all partners	Dimitra Kondyli (EKKE)
0.3	30/07/2017	3rd draft version by all partners	Dimitra Kondyli (EKKE)
0.4	21/9/2017	4th draft version by all partners	By task leader and all partners involved
1.0	25/10/2017	Final draft version submitted to CESSDA MO	By task leader

Authors List

Organisation	Name	Contact Information
EKKE	Dimitra Kondyli,	dkondyli@ekke.gr
EKKE	George Fragoulis	gfragoul@yahoo.gr
EKKE	Apostolos Linardis	alinardis@ekke.gr
CNRS	Nathalie Paton	nathalie.paton@univ-toulouse.fr
CNRS	Federico Zemborain	fzemborain@msh-paris.fr
ICS-ULisboa	Pedro Moura Ferreira	pedro.ferreira@ics.ulisboa.pt
ICS-ULisboa	Claudia Oliveira	claudia.oliveira@ics.ulisboa.pt
ICS-ULisboa	Bárbara Rodrigues	barbara.rodrigues@ics.ulisboa.pt
CSDA	Tomas Cizek	tomas.cizek@soc.cas.cz
CSDA	Jindrich Krejci	jindrich.krejci@soc.cas.cz
ADP	Janez Štebe	Janez.Stebe@fdv.uni-lj.si
ADP	Irena Vipavc Brvar	Irena.Vipavc@fdv.uni-lj.si
ADP	Irene Bolko	Irena.Bolko@fdv.uni-lj.si
SND	Martin Brandhagen	martin.brandhagen@snd.gu.se
UKDA	David Hall	djhall@essex.ac.uk
UKDA	Louise Corti	corti@essex.ac.uk
UKDA	Stuart McDonald	sm17049@essex.ac.uk
UKDA	Veerle van den Eynden	veerle@essex.ac.uk

Time Schedule before Delivery

Next Action	Deadline	Care of
Review by the task partners	(21/09/2017)	EKKE
Review by the WP leader	25/10/2017	NSD
Review by the Chair of the Delivery Committee	26/10/2017	CSDA
Review by the Project Coordinator	27/10/2017	CESSDA
Approval and Submission by the Project Coordinator to the European Commission	30/10/2017	CESSDA

Abbreviations and Acronyms

ADISP	Archives de Données Issues de la Statistique Publique
ADP	Arhiv Družboslovnih Podatkov
BDE	Big Data Europe project
CESSDA PPP	CESSDA Preparatory Phase Project
CIMES	Centralising & Integrating Metadata for European Statistics
CNRS	Centre Nationale de la Recherche Scientifique (Progedo)
ČSDA	Czech Social Science Data Archive
DA	Data Archive
DANS	Data Archiving and Networked Services
DAS	Data Archive Service
DDA	Danish National Archive - Danish Data Archive
DMP	Data Management Plan
DwB	Data without Boundaries project
EEA	European Economic Area
EKKE	Ethniko Kentro Koinonikon Erevnon
ELSTAT	Hellenic Statistical Authority
ERA	European Research Area
ESCOS	European Service Centre for Official Statistical Microdata
EU GDPR	European Union General Data Protection Regulation
EuRAN	European Remote Access Network
FORS	Swiss Foundation for Research in Social Sciences
FSD	Finnish Social Science Data Archive
GESIS	Leibniz Institute for the Social Sciences
ICS-ULisboa	Instituto de Ciencias Sociais da Universidade de Lisboa
IECM	Integrated European Census Microdata
INSEE	Institut National de la Statistique et des Etudes Economiques
LiDA	Lithuanian Data Archive for Humanities and Social Sciences
MO CESSDA	Main Office CESSDA
NSD	Norwegian Centre for Research Data
NSIs	National Statistical Offices
OA	Open Access
OECD	Organisation of Economic Co-operation & Develoopment
OS	Official Statistics
PUFs	Public Use Files
RDA	Research Data Alliance
RDM	Research Data Management
SND	Swedish National Data Archive

SNDS	Système National des Données de Santé
SORS	Statistical Office of the Republic of Slovenia
SPs	Service Providers
So.Da.Net	Greek research infrastructure for the social sciences
SUFs	Scientific Use Files
TÁRKI	TÁRKI Alapitvany (TARKI Foundation)
UGOT-SND	University of Gothenburg - Swedish National Data Service
UKDA	UK Data Archive

EXECUTIVE SUMMARY

The main objective of this deliverable is to survey the current state of play of data archiving services (e.g. amount of data available, other data providers etc.), as well as current and future researchers' needs in four domains, i.e. academia, health, official statistics and history. We also explore current and potential agreements for broadening data perimeter and relevant national and international policies. This task is also linked to the audit task carried out under T3.2. Thus, we provide an overview of each domain by summing up the types of data used and produced by researchers or data they would like to access, identifying data perimeter and potential actions for missing the gaps.

In order to provide insights concerning the aforementioned issues, we used mixed methodology, i.e a) desktop research regarding the datasets CESSDA SPs hold in the examined data domains b) interviews with experts-key informants in each domain, and c) literature review, regarding also CESSDA outcomes of previous and current research projects. At the end, we provide some examples in different domains in order to stress out best practices cases. These cases, including indicatively the collection, curation and dissemination of new types of data, or the setting up of new structures in order to collect dispersed data and facilitate researchers' access, could act as departure point for expanding the data perimeter of SPs or elaborating new strategies.

With regard to each data domain that we studied, main findings and recommendations could be summarized as follows.

Academia domain

Data of the academia domain are located at the core of CESSDA SPs activities and by definition include all kind of data (historical, health etc.) produced with different methods (quantitative, qualitative, mixed). There is an uneven distribution within SPs and CESSDA could take advantage of different SPs expertise regarding the type of data archived and managed. The challenge ahead for CESSDA consists of dealing at a greater extent with types of data that require different storage, curation and management.

An ever-larger part of the body of empirical research, beside quantitative surveys, consists of qualitative datasets of different kinds: transcripts of research interviews, but also images or audio-visual recordings etc. Compared to other researchers, the qualitative social research community is far less accustomed to the practice of secondary analysis and reuse of data sources from other researchers. Despite that, qualitative data archiving is developing dynamically, both within existing social science data archives and in the form of independent qualitative data archives. Interestingly, too, as a growing number of studies are based on a mixed-methods design, archives are more often required to ingest different types of datasets. In the field of qualitative data archiving, there is uneven development within European countries. Most SPs do not have any qualitative datasets in their catalogues, but in some of them the number of qualitative datasets is slowly rising. Major challenges for expanding use of qualitative data are the preservation, legal and ethical issues.

Thus, according to the specific needs and characteristics of each national setting, SPs should take action to:

a) Set up adequate infrastructure in order to handle with all types of quantitative data.
b) facilitate the acquisition of qualitative data collections along four lines:
    - Identify demand for qualitative data
    - Target research on specialised tools and services for qualitative data;
    - Enhancing provision of qualitative data collections by campaigning the advantages of archiving them;
    - Establish long term collaborations with other actors specialising in collecting handling qualitative data.

Health data

Nowadays, the importance of accessing and analysing medico-administrative data for the elaboration of public health policies is broadly recognized. Partially responding to this need, some SPs have recently started collecting and disseminating health data. However, this type of data clearly poses various challenges to SPs. The legal aspect should be taken into account, as health data in many cases cannot by anonymised without losing valuable information. The stake is to protect patients while enabling research. CESSDA SPs can play an important role, as the increasing need for matching health data with administrative records and public health system, as well as the difficulties in cross-border cooperation, have raised an interest for secure remote access within the research communities. Various demands are coming to light, such as the need for the development of metadata descriptions or the minimization of the variations between classification standards. SPs should elaborate strategies for enriching their health datasets by establishing new agreements with health data producers and relevant institutes/organisations outside CESSDA, as well as to develop the necessary infrastructure and skills in order to host, curate and disseminate health data.

Thus, according to the specific needs and characteristics of each national setting, SPs should establish agreements with relevant key actors and organisations outside CESSDA for hosting health data. Moreover, as health data come from different scientific fields, CESSDA could provide services such as classification standards, metadata, or documentation standards in infrastructures dealing with health issues and collaborate with governmental and other agencies involved in the development of a legal framework protecting personal information and enabling, as possible, research.

Official statistics

Whether referring to understandings of OS by the OECD or the Administrative Data Research Network, data may be drawn from all types of sources, whether statistical surveys or administrative records. Along the lines of this understanding, with the goal of widening CESSDA's data perimeter by furthering co-operation between "data archive services" (DAS) and OS, we will mainly focus on micro-data, as opposed to other types of data, namely aggregated or macro-data.

OS revolves around statistics produced by governments and boards, private corporations, or regulators providing some kind of public service. Thus, statistics are now produced by a wide

range of agencies. The increasing use of web data or transactional data, as well as the development of administrative data use are important to consider, since such evolutions raise unprecedented issues in terms of preservation, documentation and access for the social sciences, creating also new researchers' needs. Big Data is also one of the key assets of the future.

There are to be find examples of co-operation among data archives services and official statistics in all fields that have a potential for co-operation. Some of activities extend beyond individual countries. Projects with the aim to provide solution for a search portal, access to administrative register microdata, and co-operation on metadata production can be pointed as outstanding. Researchers stress the need to decrease time needed to acquire access to microdata from NSIs, as well as for more detailed documentation on methodology, high quality consistent and citable metadata, or heterogeneous metadata information throughout Europe for trans-national comparability.

Thus, the emergence of various actors, the growing number of data producers, the vaster points of access for the same dataset, the use of new devices by researchers, or yet the relationship between administrative services and local infrastructures providing DASs are a few of many examples that stress the evolutions affecting coverage. CESSDA should persist in exploring big data landscape for the social sciences by building up on current and future research projects along with the major actors dealing with big data and collaborate with other experts in order to develop methodologies and tools facilitating researchers' work.

Historical Data

History may be a peripheral discipline for CESSDA to the extent that it belongs to the humanities rather than social sciences. However, there are intersections between historical and social science research that require scholars to access data from both recent and past sources. Regarding the number of historical datasets SPs hold, there is a clear increase from the last research findings disseminated in 2012.

Some SPs provide access and disseminate a considerable amount of historical data. However, SPs should take into account the recent developments in the field and establish agreements with other institutions/organizations and data producers in order to keep up with researchers' needs, which change over time. Time series that are built on quantitative data – gathered from diverse historical sources, documents and archives are considered important and it seems more likely to be used by other social science researchers. Metadata descriptions must be developed, since historians use various historical sources in their research - from public documents to private letters - but these data were not initially collected for statistical/research purposes.

The various types of historical datasets set issues of sustainability, meaning not only to keep data alive, but also to enable the exploitation of advances in technology, as well to enable connections between resources that could lead to new discoveries and broader impact.

Thus, according to the specific needs and characteristics of each national setting, SPs should establish agreements with other key actors and institutions/organisations producing or

holding historical data, emphasise on the collection, curation and dissemination of time series and produce metadata descriptions and follow up technological advances in order to increase sustainability and allow better insights into historical datasets.

Table of Contents

1. INTRODUCTION

The overall objective of WP3 is to evaluate social science data archives and services in EEA countries to identify gaps and bottlenecks in existing data centres and services, in such a manner that national development plans can be suggested to help close the gaps and overcome present barriers in developing archiving services on a global level. Task 3.4 works towards this goal by surveying the state-of-play of DAS and researchers' needs as well as developing best practices and guidelines to support data services to widen their data perimeter.

The "tidal wave of data" evoked by major actors such as the High-Level Expert Group (HLEG) on Scientific Data of the EU, is described in a following manner:

"*A fundamental characteristic of our age is the rising tide of data – global, diverse, valuable and complex. (...) We all experience it: a rising tide of information, sweeping across our professions, our families, our globe. We create it, transmit it, store it, receive it, consume it – and then, often, reprocess it to start the cycle all over again. It gives us power unprecedented in human history to understand and control our world. But, equally, it challenges our institutions, upsets our work habits and imposes unpredictable stresses upon our lives and societies*"[1].

Fifteen years ago, Science Magazine published interviews based on all fields' scientists including SSH researchers and questioning about the future of their fields[2]. They did not foresee in 1995 the deluge of new social science data and the huge quantities of digital information to spread worldwide. CESSDA was one of the informal consortiums at that time, of which members organisations seemed to foresee the evolutions ahead and attempted in various ways to reflect upon and deal with the data deluge to come. This transformation of the landscape must be investigated further to evaluate the future directions for data archiving services. Social science and humanities data archive services must indeed consider the manners in which the emergence of new types of data bring forward unprecedented legal, ethical, technical, financial and management issues, to name just a few issues in need of attention. In the current context, there are not only new types of data but also new data sources. Private corporations, various agencies, governmental organisations, all these actors contribute to producing data anew. Some of these actors do not have the same data sharing culture as researchers, nor do they have the same prerogatives as public entities, serving citizens and democracies interests. In parallel to new types of data and an increase in the number of data sources, the points of access have multiplied. Several data providers can offer access to same data, creating confusion for users as well as improper management of resources. These changes affect directly CESSDA and CESSDA's SPs as the playing field is undergoing extensive and rapid evolutions, creating new needs and stakes that must be taken into account to prevent losing ground. For CESSDA and its partners to move ahead and plan the future at best, it is essential to have a good understanding of how data archiving services need to evolve to keep on track and serve data producers at best.

[1] European Union (2010). *Riding the wave. How Europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data.* Retrieved from https://www.fosteropenscience.eu/content/riding-wave-how-europe-can-gain-rising-tide-scientific-data

[2] Gary, K. (2011). Ensuring the data-rich future of the social sciences. *Science*, 331 (6018), pp. 719-721 http://nrs.harvard.edu/urn-3:HUL.InstRepos:12724029

Even without taking into account how the data inflation is affecting the field of archiving, it is essential to assess the current state of data production, data sharing practices as well the level of coverage of archiving services. CESSDA's future depends on its' ability to be relevant and used by researchers from SSH, which implies investigating the kind of data CESSDA SP's have, whether this data meets researchers needs, at an extent, where co-operation should be envisioned or what bottlenecks in coverage prevent the expansion or a broader visibility of services.

This task takes up these challenges; it starts reviewing the level of coverage provided by SPs and identifies some of the main challenges data archiving services are facing. Surveying the data perimeter and exploring researchers' needs for various types of data, is a first step to map the state of play of the data field. Carrying out case studies and best practice cases within CESSDA SPs is a second step to attempt providing practical roadmaps in view of contributing to the CESSDA global strategy.

The survey of the data perimeter and researchers' needs is conducted by exploring different fields of data. The Description of work suggests studying four fields in particular, i.e. historical data; academic data; health and big data; official statistics.

Using a field-based approach is a key to comprehend the changes within the European landscape over the last years. By focusing on sub-domains of data, depicting or identify existing and new kinds of data that should be included within CESSDA as well as providing useful pieces of information towards possible directions for the future. It is worth mentioning that T3.4 has connections with the aforementioned project. Therefore, it attempts going beyond in the sense that it investigates along with existing new kinds of data and tries to set up a practical roadmap based on SPS experiences and practices to meet new needs and challenges.

2. DESCRIPTION OF WORK AND ROLE OF PARTNERS

2.1 TASK 3.4: EXPANDING THE DATA PERIMETER

T3.4 puts into perspective the current perimeter of action of SPs as well as researchers' needs to provide roadmaps aiming at widening CESSDA's perimeter.

2.2 RATIONALE OF TASK 3.4

Task 3.4 is meant to strengthen and widen CESSDA scope of action by expanding the data perimeter of CESSDA and CESSDA SPs. To do so, it is essential to have a clear view of the landscape. This implies understanding upcoming challenges for archiving services as well as estimating CESSDA and CESSDA's SPs current perimeter of action. The evolution of SPs perimeter of action or the construction of partnerships between different data centres, namely NSIs and SPs, started back in the years 2010, as early as the CESSDA PPP project, and was later pursued within the DwB project. At the time of CESSDA-PPP, a series of criteria were put forward to better determine the existing data perimeter. It was suggested to:

- Identify data that is not currently accessible;
- List strategies for acquisition policies;

- Harmonise dissemination policies;
- Suggest realistic planning for networking with data agents at various phases in the life-cycle of data;
- Strengthen the bonds among research actors (i.e. academic-administrative-business actors);
- Internationalise research in terms of provision, while focusing at national/cultural boundaries in terms of production[3].

Such an enquiry implies questions such as: what is produced, how it is provided and where – if not through CESSDA? How far is social sciences research production covered by data archives from the Country members of CESSDA?[4] For some data domains, namely official statistics, much information has already been collected. In general, though, major blind spots remain and much work still needs to be conducted to determine what new formats of data are emerging, what type of datasets are outside of CESSDA's holding and why, what are the major challenges for tomorrow.

2.2.1 TWO-FOLD GOAL: INVESTIGATION OF DATA DOMAINS COVERAGE AND PATHS FOR DEVELOPMENT

To best prepare CESSDA and CESSDA's SPs for the future, T3.4 has set out specific goals, to examine the state-of-play, i.e challenges ahead & main understandings of the data domains enquired in this task and state of coverages of SPs, as well a possible path for the development of the SPs' services. In particular, we explore bottom-up definitions, issues facing data domains, researchers' needs, researchers' data sharing culture and practices, portion of each data domain that are archived, agreements and other actors outside CESSDA that affect coverage and they should be taken into account in the formulation of data acquisition policies, data sharing policies at the national and international level, as well as field related policies.

2.2.2 ORGANIZATION OF WORK

The work of academic task is allocated to the Czech Social Science Data Archive at the Institute of Sociology, Czech Academy of Sciences (CSDA) and the Institute of Social Sciences, University of Lisbon (ICS – ULISBOA). The work of health data is allocated between the Swedish National Data Service-SND, the Centre National en Recherché Sociale- CNRS and the University of Essex, UK Data Archive - UKDA, while EKKE has performed the part "proportion of data currently archived by the existing data services". The work of Official Statistics is allocated to University of lubljuana, Social Science Data Archive - ADP and to CNRS. The work of historical data is allocated to the National Centre for Social Research - EKKE and the Institute of Social Sciences, University of Lisbon - ICS- ULISBOA. Regarding the division of work among partners involved, allocation of person months has been taken into account.

[3] Report: "Data collection strategies: CESSDA organisations and their relation to data collections outside CESSDA (D10.5a)", p. 59.
[4] *Ibid.* p.59-60.

2.2.3 SCOPE OF THE INVESTIGATION: TYPE OF DATA AND COUNTRIES

As described hereinafter, the task focuses on four types of data in particular and is supposed to take into account a rather wide range of countries.

4 types of data are considered in T3.4:

1. **Academic data**: data collected by researchers within the universities/ research institutions and funded under research budgets. Both quantitative and qualitative. (CSDA & ICS-ULISBOA)
2. **Official statistics and big data**: surveys produced by National Statistical Institutes (NSIs) & administrative data. Increasing use of the survey-administrative datasets & increasing linkage with qualitative interviews. New actors providing data services. (UL-ADP & CNRS). A sub-category though of constantly increasing importance within the domain of official statistics is big data in terms of administrative-operational data as well as transaction (i.e. uses that generate data) & web data (online sources of data, (MO CESSDA).
3. **Health data**: public health surveys, epidemiological cohorts, medico-administrative data (e.g. social security data). Large perimeter to check how things are being organised. It is worth mentioning that within this particular category a growing tendency of data concerns big data. New field of investigation such as genes analysis, epidemiological studies etc provide a promising field of big data (SND, UKDA & CNRS).
4. **Historical data**: data and resources collected & used by historians/humanities. (EKKE & ICS-ULISBOA)

While these delimitations of the domains were used at the beginning of the CESSDA SaW project to divide work amongst partners, top-down definitions should be provided during the completion of T3.4 to help clarify terminologies used amongst Member country archives. Investigating online the manners in which formats of data are defined by archiving services ensure the use of bottom-up definitions, i.e. from the field/actors rather than institutions. This activity is even more relevant if one considers that CESSDA is still at an early age and can benefit from the output of this activity as CESSDA will thus be able work in the future on the basis of acting definitions rather than top-down ideas of data domains, i.e. institutional ideas of domains.

Types of Data Archives Services as part of the investigation:

- CESSDA members
- Emerging Data Services or potential other data providers (for example National Statistical Institutes)

2.2.4 READJUSTMENTS OF THE SCOPE OF THE TASK

Due to lack of sufficient time and means and in order to better focus on the remaining goals, the two following subtasks were set aside early on in the project:

- Assist the countries in their aim to widen their data perimeter concretely according to their maturity level in the domain to build agreements taking into consideration the national context;
- Involve the funding agencies and the research councils in discussions regarding systematic data policies.

Both aims are strategically important and will be an integrated part of CESSDA operations and work plans in the years to come.

Limits to identifying other actors impacting coverage & mapping datasets

Task 3.4 will examine actors impacting coverage within CESSDA' SPs as well as other actors outside CESSDA and hold datasets that could be of interest to CESSDA. These actors could be:

1. New types of actors, i.e. private corporation;
2. Other infrastructures within the public sector holding datasets potentially of interest for CESSDA and partners with which collaboration should be envisioned.

During phase 1, investigation outcomes led us mainly to the second category of actors. This task implies discerning data providers, data producers and/or data archives throughout Europe. In the field of Official Statistics for instance, several initiatives can be observed. Lists of data centres have been provided by the Official Statistics office of the UN[5] and central archives like the UKDA[6]. Lists have been generated within the frame of research conducted for EC projects, like with DwB where data fact sheets were created[7]. Identifying data providers, producers and/or archives is a manner to pinpoint the amount and types of data that is not currently accessible. In turn, it is possible to know how far CESSDA still has to go to properly cover social sciences research production.

While this step is necessary to consider data that is not currently listed through CESSDA, it is also an impossible task if the goal is to be exhaustive, because of:

- the growing number of data producers,
- the always vaster points of access for the same datasets,
- the new technological supports for generating and holding data, etc.
- for feasibility reasons and proper cumulativity, a global approach must be adopted.

Here again, the lack of sufficient time and means makes this task impossible to carry out in a fully satisfactory manner. However, an effort has been made in order to map other actors holding types of data located outside CESSDA SPs namely for historical, statistical and health data. A more exhaustive mapping of data centres that could be of interest to CESSDA can also be an integrated part of CESSDA operations and work plans in the years to come.

Big data, an all-encompassing field

Big data is an all-encompassing field, in the sense that each domain now deals with this new format of data. In the case of Official Statistics for example, administrative (e.g. government transactions) and business data can very well be considered as big data. Likewise, the huge

[5] UNstats.un.org
[6] https://www.ukdataservice.ac.uk/get-data/other-providers/data-archives/europe
[7] http://www.dwbproject.org/access/accreditation_db.html

amount of data produced in the field of health brings this latter field to be partial to big data. As for academics, big data can actually be considered as a branch in itself, part of the wider realm of academics. Historical data is also impacted since the development of new statistical software allows to process massive amounts of data.

Even if each sector seems to develop new strategies with regards to a generic idea of big data (i.e. new technical possibilities leading to new political, economical and scientific strategies), in all reality, big data mainly concerns two very specific types of data, i.e. web data and transactional data. Everything outside of that goes back to the idea that data can now be aggregated into massive sets of data, and therefore are greater in volume. To better deal with the fact big data has become an all-encompassing field each data domain studied in T3.4 can very well bring up how this field is affecting other domains, if this is relevant in the sections of upcoming challenges, knowing that big data is dealt with on its own.

2.2.5 ORGANISATION OF WORK

Outline of the final deliverable

The deliverable is designed to fit the common needs of the task without regard to the particularities of each data domain and work completed in previous projects. Adjustments may have been made for each domain but the overall outline goals can be understood as described hereinafter.

A first section sets the stage for the data field under study. It is dedicated to:

- **The description of the field under investigation**; By the means of institutional literature, the overall understandings are presented to define what the field is.
- **The goals set out for that specific field** in accordance with previous work and partners understanding of 3.4;
- **The methods used to reach the goals envisioned**.

A second section explores the main understandings of the domain as well as the new issues arising in DAS in close relationship to social sciences research production, to ensure upcoming challenges are properly integrated into future CESSDA plans and essential steps to take to ensure coverage are foreseen. It collects information on:

- **Bottom-up definitions**: Definitions presented in institutional or scientific reports are most often reliable but these approaches foster top-down definitions of domains that archives may not actually perceive as fit to the data in their holdings. Task 3.4 has the opportunity to explore bottom-up definitions, clarifying terminology used by Member country archives, observers and SaW partners (such as NSIs where collaboration is developed). The information presented here is meant to allow CESSDA to build the future based on “field definition of domains” rather than “top-down institutional definitions”.
- **Issues and stakes ahead**: For CESSDA to prepare for the future, it is necessary to consider the challenges that lie ahead and establish the main difficulties that may arise.
- **Researchers’ needs:** Likewise, it is important to consider how researchers’ needs are evolving with special emphasis to interviews conducted with the research /academic personnel and information based on each data domain.

- **Researchers' data sharing culture & practices**: And for the researchers that are not yet inclined to using national archive, CESSDA SaW is the opportunity to study research data sharing culture and practice and that prevents the expansion of archiving in general and in specific domains when appropriate. This issue, along with open access policies, has been developed in a distinct chapter (ch. 3), as similar problems and challenges to face seem to be arisen regarding all kinds of data. Thus, we only refer to field related sharing culture and practices, when appropriate.

A third section enquiries the state of play of SPs of each field and others actors affecting coverage. It is important for CESSDA to have a clear view of the landscape of data production, data provision and archiving, in order to acknowledge whether CESSDA archives are actually serving their purpose and/or if data is mainly being archived elsewhere. Information reported in this section is meant to identify data that is not currently accessible, set the ground to possibly build future strategies for acquisition policies and harmonising dissemination policies, and set the stage for realistic planning for networking with data agents at various phases in the life-cycle of data[8].

Portion of data archived & Agreements impacting coverage: This part discusses the level of coverage provided by CESSDA Country Members, i.e the datasets stored for each field by CESSDA partners, being available to the research community and/or the wider public. It is noted that the estimations provided here regarding health data and historical data, should be read as a snapshot of the specific period time taking also into account language barriers and classification issues, i.e. how datasets are classified in the SPs.

- **Data sharing policies** on international and national levels, and field related policies on a national level when is possible.
- **Other actors affecting coverage**: Considering major blind spots remain to determine what datasets are outside of CESSDA's holdings, this part is dedicated to mapping data centres and identifying new types of actors affecting DAS.

We have to mention at this point, that the above structure does not apply to the field of academic data, as most of the data stored in CESSDA SPs have been produced in universities or research institutes. Thus, issues such as researchers' needs, data sharing policies, the level of coverage provided by CESSDA SPs etc have not been included within the frame of academic data chapter. It is worth mentioning that CSDA and ICS-ULisboa institutions in charge of the academic data have significantly contributed to the development of the chapter regarding open access policies and sharing cultures, along with EKKE. Taking into consideration the broad conceptualization regarding 'academic data', we have chosen to stick to the more specific 'open academic data' and its open access policies. Thus, and bearing in mind the increase of importance of the latter within the research community, and more specifically already engaging to FP9, we believe this deliverable will contribute considerably to the advancement of knowledge on this matter.

[8] Report: "Data collection strategies: CESSDA organisations and their relation to data collections outside CESSDA (D10.5a)", p.59

3. OPEN ACCESS POLICIES

3.1 OPEN ACCESS, POSSIBILITIES AND OBSTACLES

Open Access is a broad issue covering open access to information and science in general. As Open Access constantly increases and it is located at the core of data debate nowadays, CESSDA and SPs are also concerned. Open access to research publications and research data, possibilities to link data and publications, and open research methods are said to be leading the way to instigating significant changes in scientific practices (Borg, 2014). Effective sharing of data from academic research projects depends on willingness of investigators and data producers to publish their data; it also depends on necessary resources of data management in ways that support re-uses of their data files for purposes of secondary analysis. The goodwill to do so varies, and the culture of data sharing differs between fields, research communities and countries. Therefore, science policies that promote the principles of open access to data are an important factor shaping the environment for data sharing in academic research. Of course, this is a very complicated issue, as funders have different policies and processes for maximizing the value from the datasets generated by their researchers (Expert Advisory Group on Data Access, 2015). However, it is widely accepted that where the research involves human subjects, they have moral and legal rights to protection, primarily of their privacy. Moreover, the researchers who collect the data, having invested time, effort and intellectual creativity, also have important rights (Expert Advisory Group on Data Access, 2015).
Recently, there is much discussion about the need for free access to knowledge and the establishment of alternative paths for the dissemination of scientific results, because of:
a) The limited access to scientific production.
b) The lengthy process of publication of scientific knowledge.
c) The reduced impact of research on economic development and on the improvement of the quality of life (especially for developing countries).
d) The oligopoly of academic publishers and the rising cost of the journals requiring subscription fees.
e) The difficulty of public or university libraries to pay subscription fees because of the reduced state funding.
f) The international growth of scientific production.
g) The demanding needs of users (National Documentation Centre, 2016).

The most commonly used definitions of open access are included in the Budapest Open Access Initiative (2002), in the Bethesda Statement on Open Access Publishing (2003) and in the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003) (known as the “BBB” definition of open access). In brief, in the Budapest Open Access Initiative, open access is defined as follows:

"There are many degrees and kinds of wider and easier access to this literature. By 'open access' to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The

only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited".

In the Bethesda Statement and in the Berlin declaration, open access is approached as follows:

> *"For a work to be OA, the copyright holder must consent in advance to let users "copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship...."* (Suber, 2015).

3.2 OPEN ACCESS AT THE INTERNATIONAL LEVEL

To a large extent, international and European policies set the stage to the framework for the development of related national policies. Besides different interest groups, the key players in formation of the open access policies regarding academic data are the OECD and the European Union (EU).

In 2004, an intergovernmental declaration (OECD, 2004) was adopted by the OECD and signed by 34 countries, with a commitment to work towards the establishment of access regimes for digital research data from public funding in accordance with defined basic principles including openness, transparency, legal conformity, professionalism, protection of intellectual property, interoperability, efficiency, and accountability. The European Union joined this declaration as well. The OECD declaration was signed by 23 CESSDA SAW target countries; eight of the remaining states are EU members; and the rest (Belarus, Kosovo, Serbia and Ukraine) do not have any commitment. While the declaration does not commit to any concrete measures, at least it has led in the signatory countries to setting the necessary agenda and raising awareness of the problem.

A more specific document on "*OECD Principles and Guidelines for Access to Research Data from Public Funding*" was published (OECD, 2007) in relation to the declaration. OECD noted that various formats of data produced within the research realm, such as administrative data, data from health organisations, geo-spatial data or scientific databases are increasingly used beyond the original project for which they were gathered. It also stressed out that data produced in publicly funded projects and for the purposes of public research should be made available for re-use in research, unless there are serious reasons preventing publication of this data. Thus, OECD shaped a list of principles and guidelines on access practices that should apply to conduct research on data produced by the means of grants coming from public funding. The principles are not applicable to commercial research or research for commercialization purposes. An exception has been granted to data that cannot be published due to personal data protection, security or other justified concerns as defined in the Principles. The Principles are enshrined also in the goals and data policies of CESSDA.

The European Commission has been exerting a more considerable pressure on the introduction of open access. The Digital Agenda presented by the European Commission (EC) forms one of the seven pillars of the Europe 2020 Strategy project (European Commission, 2010). The central aim of the EU 2020 strategy is to put Europe's economies onto a high and sustainable growth path. Europe has to use its resources in the best way possible and one these resources

is public data, meaning all the information that public bodies in the EU produce, collect or pay for. To this end, EC proposed a package of measures to overcome existing barriers and fragmentation across the EU, as part of the Digital Agenda for Europe. The following measures are meant to work towards these goals:

- *Adapting the legal framework for data re-use. A proposal for a revised Directive on the re-use of public sector information and a revised Commission Decision on the re-use of its own information are adopted together with this Communication*,

- *Mobilizing financing instruments in support of open data, and deployment actions such as the creation of European data-portals*,

- *Facilitating co-ordination and experience sharing across the Member States. According to the EC, open data are expected to increase economic opportunities, address societal challenges and accelerate scientific progress* (European Commission, 2011).

While OECD policy is defined specifically for research data sharing, the EU pursues access to research data primarily as part of more general agenda of open access to scientific information (see e.g., European Commission, 2007; 2012a; 2012b; 2013, Council of the EU 2016). In this way, the EU strives for open, collaborative research environment based on reciprocity, which will be ensured in the whole European Research Area (ERA). In the context of providing access to peer-reviewed scientific publications, data are perceived as part of such publications and the access to data should be granted together with this publication to allow verification of published analysis (European Commission, 2012b). In addition, research data comprise a special part of the open access agenda wherein emphasis is placed not only on transparency of research but also on data re-use.

However, accessibility of data is not a single condition allowing secondary data usage. Therefore, the key points of EU open research data strategies include also building of appropriate research infrastructures that are not limited to long-term storage, but also serve to ensure the more general conditions for data re-use and promotion to projects improving the practice of data management, data awareness and the data sharing culture in general (e.g, European Commission, 2013). Since 2014, the European Commission is implementing the open research data pilot in the Horizon 2020 Framework Programme (see European Commission, 2016a: Art 29.3; 2016b).

The establishment of the Open Science Policy Platform witnesses the determination of the EC to promote data openness (see also Council of the European Union 2015; European Commission, 2015 for the results of the public consultation on Science). The Council of the European Union (2016) underlines that

"Research data originating from publicly funded research projects could be considered as a public good… whilst recognising simultaneously the needs for different access regimes because of Intellectual Property Rights, personal data protection and confidentiality, security concerns, as well as global economic competitiveness and other legitimate interests. The underlying principle for the optimal use is summed up as follows 'as open as possible, as closed as necessary."

The Council of EU (2016) has welcomed the intention of the Commission to make research data produced by the Horizon 2020 programme open by default (taking under consideration the previously mentioned possible restrictions), to make the cost for preparation and management of research data eligible for funding. It also encourages Member States and stakeholders to implement Data Management Plans as an integral part of the research process. Finally, it emphasises the importance of storage, long term preservation and curation of research data, as well as ensuring the existence of metadata based on international standards. While its recommendations have been assigned different priority ranks in different member states, the staged introduction of open research data pilot under the Horizon 2020 has importantly stimulated the defining of adequate data policies at the national level. According to the EC recommendations the EU Member States should enforce the same principles for national research funding. Pilot projects in several Horizon 2020 areas started in 2014 and since 2016 the pilot has expanded to the whole programme. Thus, any neglect of open access issues may lead to concrete consequences for countries' competitiveness in international scientific collaboration in the ERA.

3.3 OPEN ACCESS AMONGST CESSDA SAW TARGET COUNTRIES

While general principles and policies on open access are defined at the international level, their implementation into practice takes place at the level of research funding agencies. The specifics of different fields and subsidy programmes are reflected in funders' more detailed data policies. For example, data management plan is required as compulsory part of each grant application in Horizon 2020 and its compliance with the data policy is assessed during evaluation process. Consequently, the data management plan is included into the grant agreement and its realization is monitored and enforced. Such data policies allow to follow the rule "*as open as possible, as closed as necessary* "(e.g. Horizon 2020) considering both, (1) open research data principles and (2) wide diversity of research environments.

There are considerable differences in the state of open access among the CESSDA SaW target countries. A recent EC study on the implementation of existing recommendations (European Commission, 2016c) divided the countries into the following three categories:

1. Very little or no open access to research data policies in place and no plan for a more developed policy in the near future.
2. Very little or no open access to research data policies in place, but some plans in place or under development.
3. Open access policies/institutional strategies or subject-based initiatives for research data already in place.

Table 1 shows the categorisation of some CESSDA SaW target countries. Others are ignored by the study because they are neither EU member states, nor associated countries. Yet other countries were not classified because they failed to supply the underlying information.

Table 1: Classification of CESSDA SaW target countries into three groups according to the existence of policies on open access to data

Group 1. Very little or no OA data policies in place / no plan
Cyprus, Latvia, Luxembourg, Malta, Poland
Group 2. Very little or no OA data policies in place / some plans
Austria, Belgium, Croatia, Czech Republic, Estonia, Hungary, Italy, Portugal, Romania, Slovakia, Sweden, Turkey
Group 3. OA data policies/strategies/initiatives already in place
Denmark, Finland, France, Germany, Ireland, Lithuania, the Netherlands, Norway, Slovenia, the United Kingdom
CESSDA SaW target countries not included in the Study
Albania, Belarus, Bosnia and Herzegovina, Faroe Islands, FYROM, Israel, Kosovo, Moldova, Montenegro, Russia, Serbia, Ukraine
CESSDA SaW target countries not classified
Bulgaria, Greece, Iceland, Spain, Switzerland

Source: European Commission (2016c)

Table 1 makes it apparent that effective implementation of open access policies is by no means the rule. Most countries (Groups 1 and 2) are only at the start of the process, while the most advanced Group 3 is highly diversified. Although the study acknowledges that some data policies are already in place in Group 3 countries, it fails to determine the progress of their implementation. The United Kingdom has achieved outstanding results through its proactive approach to the implementation of open access policies, thus outpacing European Commission policies. Number of measures at provider level has been in place for several years. UK's data policies define not only the conditions but also the methods of data publication, including the concrete infrastructure responsible for availability (see DCC 2009–2016). A data policy of the Economic and Social Research Council (ESRC) applicable to its grant programme has been in place since 2010 (see ESRC 2015 for current wording). As result, successful applicants must provide all time access to their data in prescribed ways, unless reasons foreseen by the policy prevent them from doing so. Other Group 3 countries are actively pursuing the formulation of their data policies but in practice, they have only implemented partial measures thus far. For example, in Germany, there is no official national strategy, although data policies have been promoted by important stakeholders including research funding agencies. Also, in 2015, Slovenia approved a pilot project of open access to data which is analogical to Horizon 2020 in 2014 -2015 (social sciences are included in the pilot), while concrete measures are still being prepared. Although that it could be said that the environment is changing, important barriers to access to research data are still prevalent in the academic environment (e.g., European Commission 2013; 2016c) and a large part of data remains unavailable (see e.g., Tenopir et al., 2011).

3.4 RESEARCHERS' SHARING CULTURE

Although important, the establishment of open access policies is not enough. As Falt (2015) argues, "open access has become a mantra that is repeated over and over again in scientific research discourse, and its benefits are recognised by many. It does, after all, add to the openness of research". However, she wonders if this is enough in terms of development and progress in the field and she poses the question whether open science can work if there are problems in data sharing. She concludes that "the triumph praising open science may remain too ambiguous or, at worst, mere talk, if open access is not tied more firmly to researchers' own reality".

There are important differences between European countries: the archives differ by number of datasets, which is a result of the shape of social sciences in the country, of the extent to which existing and preservation-ready datasets are received by archives, and the proportion of data that remains unpreserved and unavailable for reuse. This tends to be a high proportion in countries where archives are absent or extremely small. In such cases the archiving of existing data is the primary challenge and problem.
Our recent survey by CESSDA SaW conducted in 2016 regarding the period 2011-2016 showed that of the 43 surveyed European countries, only 11 reported that there was a positive approach to the data sharing among the scientific community, and another 13 countries responded that there is a neutral or mixed attitude for data sharing.

Table 2: Attitudes to data sharing in scientific communities

	Frequency	Percent
Mainly negative attitudes	6	14
Neutral or mixed attitudes	13	30
Mainly positive attitudes	11	26
Unable to provide estimate	13	30
Total	43	100

If we look at the more detailed description of the status of data sharing in different countries, the results are as follows: only 7 countries have reported that they share data at least at medium and higher levels (but we can assume that data sharing in some of these countries is rather through informal channels than using the data archives).

Table 3: Sharing and access to research data for reuse

	Frequency	Percent	Countries
Rare: proportion of researchers sharing and having access to data low	11	26	Bosnia and Herzegovina, Bulgaria, Hungary, Israel, Kosovo, Lithuania, Portugal, Romania, Russia, Serbia,
Not that common: proportion of researchers sharing and having access	10	23	Belgium, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Netherlands, Slovakia, Slovenia, Switzerland

to research data low or medium			
Very common: proportion of researchers sharing and having access to data medium or high	7	16	Germany, Macedonia (?), Montenegro (?), Norway, Poland, United Kingdom
Unable to provide estimate for 2011-2016	5	12	
Missing	15	35	
Total	43	100	

Respondents of our survey tried to estimate the extent in which data in particular countries are shared and reused in the period between 2011 and 2016.

Table 4: Proportion of social science researchers that have shared the research data they produced between 2011 and 2016 (estimate based on experience by institution and publications).

	Frequency	Percent	Countries
low (0-10%)	14		Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Finland, Hungary, Kosovo, Latvia, Portugal, Russia, Serbia, Slovakia, Slovenia, Switzerland
medium (10-30%)	8		Czech Republic, Denmark, Estonia, Lithuania, Macedonia (FYRM), Montenegro, Norway, Poland
high (>30%)	2		Germany, United Kingdom
Unable to provide estimate for 2011-2016	9		Albania, Belgium, Greece, Ireland, Israel, Italy, Netherlands, Romania, Sweden
Total	33		

Table 5: Proportion of social science researchers per country able to access existing third-party data between 2011 and 2016 (estimate based on experience in given institution)

0	Frequency	Percent	Countries
low (0-10%)	8		Bulgaria, Cyprus, Denmark, Hungary, Kosovo, Latvia, Romania, Serbia
medium (10-30%)	6		Croatia, Finland, Lithuania, Montenegro, Slovenia, Switzerland
high (>30%)	8		Belgium, Germany, Macedonia, Netherlands, Norway, Poland, Slovakia, United Kingdom
Unable to provide estimate for 2011-2016	11		Albania, Bosnia and Herzegovina, Czech Republic, Estonia, Greece, Ireland, Israel, Italy Portugal, Russia, Sweden
Total	33		

So, it seems data sharing is far from being a common practice in many European countries. Hereinafter, we will have a closer look at researchers' sharing culture in the target countries

based on SAW survey. More information regarding sharing culture in specific kind of datasets will be provided in the respective chapters.

In France, the question about data sharing culture in the country was not filled in. It seems that there are efforts supported by the Ministry of Education and major public research organisations such as the CNRS to establish a data sharing culture that could be extended to more comprehensive policies and guidelines on a national level.

In Switzerland, data sharing culture is underdeveloped and the proportion of researchers sharing data in the examined period is estimated as low (0-10%). The proportion of researchers able to access existing third-party data they need is estimated as medium (10-30%). There are no available statistics regarding data sharing channels and routines in Switzerland. As the self-assessment results shows, the attitudes tend to be mixed. Even though social science researchers in Switzerland in general seem to be worried about data misuse and misinterpretation, and consider data sharing costly and time consuming, they acknowledge that there are some benefits to data sharing. There are no career rewards - related to data sharing.

In Hungary, data sharing and reuse are low, as the proportion of researchers sharing data and the proportion of researchers able to access existing third-party data they need are both estimated as low (0-10%). The most usual way to share data is via informal channels - through project or personal websites. Less preferred methods include data shared as supplementary data in a journal and archiving in repositories. There are no career rewards - related to data sharing.

In Greece, no estimation can be made regarding data sharing culture in Greece, as there are no available data. Based on respondents' experience and some relative reports, is seems that data sharing and reuse is not that common in Greece. The proportion of researchers sharing data in the examined period is estimated as low (10%). Most popular data sharing channels include formal and transparent channels, as data archive and repositories are ranked first. Data is shared also via personal contacts (ranked second) and project or personal websites (third), that lack formality and transparency.

In Czech Republic, there are no relative data, thus, estimations were based on CSDA experience. The conclusion is that attitudes towards data sharing differ between institutions, as well as between disciplines. For example, in some disciplines, such as demography, psychology, social geography, sharing culture is limited. However, this estimation maybe is due to the lack of data archives for these disciplines and probably researchers share their data via informal routes. On the other hands, in some university departments, such as the departments of sociology or political sciences, a large part of the researchers shares their data. No career rewards for data sharing have been established.

In Finland, data sharing and reuse among is not so common. The proportion of researchers sharing data in the examined period is estimated as low (0-10%), while the proportion of researchers able to access existing third-party data they need, is estimated as medium (10-30%). Most popular data sharing channels includes informal contacts (peers and colleagues)

(ranked first) and then data archive (ranked second). There are no career rewards related to data sharing.

In Germany, data sharing and reuse is very common. The proportion of researchers sharing data and the proportion of researchers able to access existing third-party data they need in the examined period are both estimated as high (above 30%). According to the researcher, in social sciences sector there is more awareness on this topic than in the natural or medical sciences, because in social sciences there are existing infrastructures. The most popular data sharing channels include formal and transparent channels, with data archives or repositories ranked first and journals ranked second. Data are also shared via projects or personal websites (ranked third) and personal contacts (ranked fourth), channels that lack formality and transparency. The attitudes of researchers toward data sharing in the social science community can be characterised as positive, even if they think that there is a risk of data misuse and misinterpretation. Currently, there are no career rewards related to data sharing.

In Lithuania, the proportion of researchers sharing data and the proportion of researchers able to access existing third-party data they need are both estimated as on medium level (10-30%). This could be explained by an initiative of Research Council required to deposit the research data at the Lithuanian Social Science Data Archive for few years. Most often used data sharing channel is formal, as data archive or repository is ranked first. Data is shared also via personal contacts (ranked second), that lack formality and transparency. The attitudes of researchers towards data sharing can be characterised as neutral or positive. However, there are some concerns regarding competition in science and publication opportunities as well as misinterpretation or misuse of data. There are no career rewards related to data sharing.

In Norway, data sharing and reuse are estimated as very common –based on the number of projects funded by the Research Council of Norway (and thus bound to share data). The proportion of researchers sharing data is estimated as medium (10-30%), while the proportion of researchers able to access existing third-party data they need is estimated as high (above 30%). As there are RDM policy requirements, most often used data sharing channel is formal and transparent - data archive or repository is ranked first. Data is shared also via personal contacts (ranked second) and project or personal websites (third), that lack formality and transparency. The attitudes of researchers towards data sharing can be characterised as neutral or negative. Even though social sciences researchers acknowledge general benefits of data sharing, there are concerns regarding negative competition as well as misuse of data. There are no career rewards related to data sharing.

In Sweden, SND staff could not provide with estimations about the proportion of researchers sharing data or being able to access existing third-party data they need. Similarly, they were not able to rank routines for data sharing.

In the Netherlands, DANS staff could not provide with estimations about the proportion of researchers sharing data. However, a difference in sharing research data among disciplines is noted. Scholars in quantitative Political Sciences and Sociology are more accustomed to sharing and use research data of others than in other disciplines. Some researchers only make use of survey data collected by international survey programmes, while others, often more qualitative oriented, collect their own data. The proportion of researchers able to access

existing third-party data is estimated as high (>30%). It can be explained by the fact that DANS is well known among the scholars in the Social Sciences and the Humanities, while researchers can also make use of the data of Statistics Netherlands (CBS). Most popular data sharing channels include formal and transparent channels, with data archive ranked first. Data is also shared via informal contacts (ranked second), project websites (ranked third) and supplementary data in a journal (ranked fourth). There are some indirect career rewards related to data sharing, as there are researchers whose articles are more cited when the associated data are publicly available.

In UK, it is estimated that more than 30% of the social science researchers have shared research data in the examined period. Concerning the use of shared social science data, it is estimated that more than 30% of the social science researchers have accessed data from the UK Data Service. A real ranking of the preferred routines of data sharing for the social science research community was difficult to provide. According to the respondents' impression, the risk that others may misuse and misinterpret data is still a thought holding back social science researchers from sharing their data, as well as the fear of time consuming effort and high costs.

By concluding this chapter, we stress the importance of funding as an incentive for enhancing data sharing and gradually developing a data sharing culture. However, drawing from SAW survey, in only three countries Data Management Plans are required in applying for a scientific grant. In other countries, this activity is not even formally required.

Table 6: Requirements or recommendations about Data Management Plans (DMPs) as integral part of on-going project activity per country

	Frequency	Percent	Countries
None	8		Bosnia and Herzegovina, Cyprus, Czech Republic, Greece, Ireland, Macedonia, Romania, Slovakia
Initial: There is growing recognition and awareness of need to require DMP	16		Albania, Belgium, Bulgaria, Croatia, Denmark, Estonia, Germany, Israel, Italy, Kosovo, Latvia, Lithuania, Montenegro, Poland, Serbia, Sweden
Partial: There is the expectation or recommendation to have DMP in place	6		Hungary, Norway, Portugal, Russia, Slovenia, Switzerland
Defined: Formal requirement, little monitoring and support	2		Finland, Netherlands,
Managed: DMP is a requirement, clear guidance is issued, support and tools are provided, the content of DMP and exemption	1		United Kingdom
Total	33		

Public research funding organisations in most European countries haven't issued requirements or recommendations about quality-assured social science research data with associated metadata, only three countries stated that these requirements are formally defined (Netherlands, Finland, Switzerland) and in only two countries are these fulfilled, including sanctions for non-compliance (Norway, United Kingdom).

4. ACADEMIC DATA

4.1. SUBTASK DESCRIPTION

The goal of our work is to describe the main understandings and upcoming challenges in the realm of academic data as well as examine the current state of preservation of data from social science research studies that are undertaken in the academic sector to outline the possibilities of preserving, storing, sharing new types of data in future. Given that academic data are located at the core of CESSDA SPs activities and by definition include all kind of data (historical, health etc.) produced with different methods (quantitative, qualitative, mixed), issues related to coverage, researchers' needs or sharing cultures are not dealt within this chapter.

4.2 DEFINITION OF THE DOMAIN

For our purpose, academic data are social science data collected by researchers within universities or research institutions and usually public funded, both quantitative and qualitative. In any case a clearer distinction may arise accordingly to the producer of data ie. the academic sector on the one hand and the public or commercial sector on the other hand.

The definition below is inspired by the OECD Guidelines for Access to Research Data from Public Funding (OECD, 2007) and defines academic data with respect to different types of data, but it is limited to social science[9] data developed for research purposes in the academic sector.

> *"Social science research data" can be defined as factual records (numerical scores, textual records, images and sounds) used and developed as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings. A research data set constitutes a systematic, partial representation of the subject being investigated."*

4.3. GOALS AND PHASES OF THE SUBTASK

Our task will consist of several stages. First, we will undertake an analysis of existing policies and strategic documents on issues of preservation and access to data, on one hand, and of the specific state of preservation of social science data in the different countries of Europe. We will build on the analysis in by conducting a more in-depth analysis based on our survey.

4.4. METHODOLOGY

We gathered information by means of an extensive review of available reports and documents of relevance to European countries' policies on data preservation and access to data, reports on the state of preservation of social science data in European countries, and final outputs of concluded projects that addressed the topic in one way or another. Our job also included an examination of the websites of existing archives, both within and outside CESSDA to: (1) explore definition of academic data (2) examine what types of data are currently preserved in

[9] Original definition deals with all the science data not only social science data

data archives. We are going to enhance our findings by incorporating the results from Cessda survey within European research area.

4.5 SCOPE OF THE DATA DOMAIN: MAIN UNDERSTANDINGS & EMERGING ISSUES

4.5.1 TRENDS AND STAKES IN RESEARCH

4.5.1.1 SHARING RESEARCH DATA: IMPLICATIONS FOR THE DEVELOPMENT OF SOCIAL RESEARCH

The practice of sharing research data has fundamental implications for the development of contemporary social research in the academia, as it has been developed in detail in chapter 3. There are at least 6 ways of sharing data: private management, collaborative sharing, peer exchange, transparent governance, community sharing and public sharing (Van den Eynden & Bishop, 2014, p. 22). Making data from research projects available for secondary analysis through public sharing expands the opportunities for combining diverse data sources, thus considerably broadening the horizons of scientific inquiry. In particular, opportunities for comparison between countries and in time are greatly enhanced. Furthermore, the data and documentation thereof provides a basis for doing methodological research, testing research instruments and designing new projects. Open access to data supports the verification of results and transparency of science.

Thus, the production of data entails considerable public expenditure, and to expand the opportunities for using the data is the logical step towards more effective research spending. Moreover, data is more than a passive source for research; access to data transforms the methodology and organisation of scientific work. Secondary analysis is becoming increasingly relevant, which affects the procedures in use. Finally, access to data is often frequent precondition of research competitiveness and involvement in international collaboration projects.

Suitable environment for the broadest and most effective sharing of academic research data possible is only guaranteed by the specialised research infrastructure of data archives and specialised projects (data services based on long-term research projects, data inventor inventories, data information systems). The research infrastructures in European countries are highly diverse in terms of their founders, legal forms, number of staff and, finally, amount and type of datasets preserved.

Survey data collected by academic researchers have driven the activities of social science data archives for a long time. Despite the vast number of datasets provided by CESSDA archives (around 25,000 in 2012), the emergence of new types of data (opinion polls, NSI data, health data, etc.) and data producers (NSI, governments, banks, etc.) has led to the understanding that CESSDA archives need to diversify their content soon (Kondyli et al., 2012, p. 5).

In 2012, politics was an over-represented subject in most archives, while less than 50% had collections concerning history, information and communication, transport, travel and mobility, etc. Subjects that remained outside CESSDA, i.e. were mostly covered by non-CESSDA organisations, were the following: economics; trade, industry and markets; education; housing

and land use planning; natural environment; law, crime and legal systems; trade, industry and markets; natural environment (Kondyli *et al*, 2012). In addition, new types of data and data producers emerged and reshaped the data landscape.

In social science research, the importance of data sources other than sample surveys has been growing over the past couple of decades with the growing importance of qualitative data, big data and government microdata.

4.5.1.2 THE EXPANDING USE OF QUALITATIVE DATA: PRESERVATION, LEGAL AND ETHICAL CHALLENGES

An ever-larger part of the body of empirical research, beside quantitative surveys, consists of qualitative results. There are different kinds of qualitative datasets – typically transcripts of research interviews, but also images or audio-visual recordings. Compared to other researchers, the qualitative social research community is far less accustomed to the practice of secondary analysis and reuse of data sources from other researchers. Despite that, qualitative data archiving is developing dynamically, both within existing social science data archives and in the form of independent qualitative data archives. Interestingly, too, as a growing number of studies are based on a mixed-methods design, archives are more often required to ingest different types of datasets. In the field of qualitative data archiving there is uneven development within European countries. Most of the archives do not have any qualitative datasets in their catalogues. But several archives started to archive this type of data and the number of qualitative datasets is slowly but steadily rising.

Table 7: Qualitative data holding in various European countries

Name	UK Data Service	Finnish Data Service	GESIS	QualiService, Bremen	Slovenian Data Archive	Swiss Data Service
Start	1994 -	2003 -	2010 -	2000 -	2004 -	2010 -
Number of datasets	1027	177	64	16	16	10

Source: Louise Corti, UK Data Service, Data Impact blog, 2016: A year of great progress in qualitative data archiving and exchange http://blog.ukdataservice.ac.uk/2016-a-year-of-great-progress-in-qualitative-data-archiving-and-exchange/

Preservation of this type of data may place different demands on the archive's technical infrastructure, compared to sample surveys and, more generally, quantitative statistical data. However, a much greater challenge to the preservation effort is posed by ethical and legal issues related to the protection of research participants' personal data. Qualitative data is much more difficult to anonymise than survey data, and in the absence of anonymisation, informed consent must be obtained from the respondents.

4.5.1.3 BIG DATA

Academic research increasingly relies on extensive data from different databases, social networking sites and such, collectively referred to as "big data". This term is currently in fashion as leading social scientists are talking about big data and analysis thereof as a fundamental

change (Savage & Burrows, 2007), or even revolution, in the landscape of empirical social science. However, the question of archiving this type of datasets remains open, oftentimes because they are not produced by academic researchers or public bodies but rather corporations such as Facebook, Google, multinationals or mobile network operators. Their considerable economic potential is primarily related to analyzing consumer behavior, advertising effectiveness and the impact of marketing activities. This poses important obstacles to preservation of big data by existing archives, which would have to address technological issues of large-volume data storage, legal issues related to personal data protection, and finally, the very willingness of the above-mentioned producers to share their data for academic reuse. The importance of this type of data is going to grow, just like the problems with sharing it, and already there are initiatives[10] and projects underway[11] that investigate the possibilities of utilizing this type of data.

There are already some attempts in this area, GESIS and UK Data archived datasets from Twitter social network but there are big limitations in archiving these types of datasets. Tweets themselves are not archived, only links to them. Tweets are then retrievable trough twitter user interface (API). If the tweet is deleted from social network, link is no longer functional. Ownership of the social media content, legal and ethical questions connected with this type of data is also of big concern.

4.5.1.4 GOVERNMENT MICRODATA

Another type of data, whose importance is growing significantly are government microdata. Microdata can be defined as "*data on the characteristics of units of a population, such as individuals, households, or establishments, collected by a census, survey, or experiment*" (OECD, 2005). Government microdata includes data collected by the National Statistical Institutes (NSI) and the administrative microdata collected by the governments themselves (Tubaro *et al*, 2012, p. 1). These data sources are considered to "*provide valuable bases for longitudinal analysis and for public policy evaluation*" (*Ibid.*: 3). Concerning their use, "*a tendency to replace surveys by administrative data and/or to merge survey and administrative data in order to reduce respondent burden, has substantially enriched these data sources and has made them even more attractive for social scientists*" (*Ibid.*). Some examples of administrative data are social security payment records, educational attainment records, health records, court records, tax records (UKDA, 2016).

In line with policies and push for open access, the use of government data is seen as a trend that will continue to grow. In this sense, previous reports stated that CESSDA archives should "rethink their policies concerning their role as intermediaries for government microdata" (Kondyli *et al*, 201, p. 37). Moreover, governments are increasingly disseminating their own data online. For instance, https://data.gov.uk/ provides 36,290 datasets covering topics such as business and economy, environment, crime and justice, defence, education, health, etc.

Research initiatives such as the Data without Boundaries (DwB) project played an important role in developing the field and strengthening relations between CESSDA and the European

[10] https://www.big-data-europe.eu/about/
[11] http://seriss.eu/about-seriss/work-packages/wp6-new-forms-of-data-legal-ethical-and-quality-issues/

Statistical Systems (ESS)[12]. DwB was a 4-year project that intended to overcome barriers across Europe concerning access to official microdata that is outside CESSDA collections considering its high value for research activities. The project tackled legal frameworks, procedures and technologies in order to access very detailed microdata in a secure way through remote access[13].

Several important outputs were produced in the framework of DwB, the most important of which was a model for an integrated service centre for the European Research Area: The European Remote Access Network (EuRAN). EuRAN is "a network infrastructure characterised by high security standards, interoperability with existing infrastructures, support for different means of access depending on differing security needs, and a single point of access as service hub for a wide range of tools and services" (DwB, 2016). EuRAN would be managed under a larger service, the European Service Centre for Official Statistical Microdata (ESCOS), which would function as a service-unit of the future CESSDA-ERIC (Silberman, 2012a). Pilots for the EuRAN brought together 3 countries each one of them hosting a Research Data Centre (RDC) and working in partnership with one CESSDA Service Provider (SP).

- CBS – Statistics Netherlands (CESSDA SP: DANS)
- Centre D'Accès Securisé aux Données (CESSDA SP: PROGEDO)
- Secure Data System (CESSDA SP: UKDA)

RDC are facilities in charge of providing access to highly sensitive and detailed data – such as administrative data, social survey, census and business microdata – through secure data systems (Silberman, 2012b). Examples of data archived in RDC are detailed geography, industry, occupation, health and demographic variables (Afkhami, 2013). RDCs are one of the strategies set by statistical authorities "to ensure privacy for individuals and to serve the needs of the scientific community" (Bender & Heining, 2011, p. 10). Research projects such as SHARE and the Luxembourg Income Survey are using the same facilities (Kondyli *et al*, 2012). Due to confidential issues, government microdata requires technical solutions that go beyond the facilities of CESSDA Service Providers.

Some CESSDA SPs have agreements in place with Research Data Centre, allowing their users to access government data. The extent to which SP are connected to the RDC can vary. The subject of the data made available through RDCs mostly corresponds to subjects that remain outside the traditional academic data perimeter. This allows us to conclude that agreements do affect the data coverage.
Further to DwB a number of CESSDA SPs has been involved in the implementation of SERISS project (CESSDA MO, NSD as leader of the WP6 on Legal, ethical and quality issues and UKDA

[12] The European Statistical System (ESS) "is the partnership between the Community statistical authority, which is the Commission (Eurostat), and the national statistical institutes (NSIs) and other national authorities responsible in each Member State for the development, production and dissemination of European statistics" http://ec.europa.eu/eurostat/web/european-statistical-system.

[13] There are different modes of access to highly sensitive data but remote access is considered to fit better researcher's needs since the researcher can be connected to the data centre from his home institution (Tubaro *et al*, 2011).

as leader of T6.3) aiming at the preparation of CESSDA SPs to handle new types of data, with a particular focus on social surveys and the use of new data types in a social survey context, including biomarker, social media data and administrative data (SERISS,2017).

4.6. EXPANDING DATA PRODUCTION AND THE PRACTICE OF ARCHIVING IN SOCIAL SCIENCES

There is a considerable difference between European countries in data production itself. The outcomes of the SAW survey conducted in 2016 regarding data production are presented in Table 8 (below).

Table 8: Characterisation of the average production of research data by the social science institutions per country

	Frequency	Percent	Countries
Rare production (data are produced ad hoc)	3	10	Bosnia and Herzegovina, Kosovo, Romania,
Periodical production (institutions have tradition in producing some type of research data to a certain extent)	17	50	Albania, Belgium, Bulgaria, Croatia, Cyprus, Greece, Hungary, Ireland, Israel, Italy, Lithuania, Macedonia, Montenegro, Poland, Serbia, Slovakia, Slovenia
Frequent production (institutions have well established tradition in data production)	13	40	Czech Republic, Denmark, Estonia, Finland, Germany, Netherlands, Latvia, Norway, Portugal, Russia, Sweden, Switzerland, United Kingdom
Total	33	100	

Preservation of data in archives is determined to a large extent by the methods of scientific work applied in the different countries, as well as by local laws and science policies. It is noted that there are considerable differences between current data access policies, with no harmonisation at the European level. In the absence of a database of data outputs from the different research projects, it is very difficult to estimate the extent to which the data production in each country is preserved by archives within and outside CESSDA. This situation might change significantly if the scientific practices changed and if the practice of dataset citation was introduced in addition to the usual publication citations.

Nowadays, the volume of research data shared is increasing but implementation of these standards is still underway. For a long time now, it has been necessary to provide for this practice in scientific codes of ethics so that all institutions and individuals who participate in the production of scientific knowledge are acknowledged fairly. In 2014, a team of more than 40 experts from 25 different organisations followed up on the Amsterdam Manifesto[14] and defined a general set of Data Citation Principles[15]. Thomson Reuters established a new Data

[14] https://www.force11.org/amsterdam-manifesto
[15] https://www.force11.org/group/joint-declaration-data-citation-principles-final

Citation Index[16], a quantitative metric which covers data sources analogously to the coverage of scientific journals by Web of Science. These activities have only just been launched, but if they are successful, they will substantially improve the situation of accounting for, archiving and reuse of social science data.

Another challenge is to add data collections from social science disciplines other than sociology and political science that are either not at all archived in data archives or comprise only a small part of their data collections. As for social science data, this is especially the case of psychology. In contemporary psychology and beyond[17], there is an ongoing debate about a so-called reproducibility crisis, namely that repeated measurements often fail to confirm the results published in scientific journals. Lack of access to high-quality primary data is one of several reasons behind this situation. This poses a relatively big problem because verification of scientific procedures is one of the fundamental pillars of modern science. Data archives are prepared to address the problem of insufficient access to data by making their existing infrastructures available for the preservation of more data. However, motivation of the scientific community in each discipline is also a necessary condition of the solution.

[16] http://thomsonreuters.com/en/products-services/scholarly-scientific-research/scholarly-search-and-discovery/data-citation-index.html

[17] http://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970?WT.mc_id=SFB_NNEWS_1508_RHBox

5. HEALTH DATA

5.1 SUBTASK DESCRIPTION

Health data is not a new field but a field undergoing changes that need to be better understood for the future of CESSDA. Health data is a large data domain that consists of a variety of study types and different methodology epidemiological cohorts, national and international surveys, medical administrative data, which could be social security data, quality registers, and public health surveys. Health data could also be divided in statistical data relative to general public health indicators, data relative to external, environmental, non-personal factors, data relative to the contextualisation of an individual's health and her psycho-social integration, genetics, and indicators relative to public health strategies. The changes in the field of health data concern the emergence, in parallel to the data archives for social science, of data infrastructures in health i.e. dedicated metadata portals, secure access systems for highly confidential and sensitive data, with similar standards, sharing issues, privacy protection etc.

The changes also concern an increasing need to link health data with socioeconomic data. This need was partially addressed through population health surveys, which are used on these topics at least since the 1970/1980 decades. Usually, the infrastructures for health data are different from those for the social sciences, but in some cases, connections between them seem promising.

The idea behind the investigation of the Health data-domain is to determine how the data infrastructures in this field is currently being organised given the recent changes in the field, what are the national policies in this domain particularly regarding links with the data infrastructures for the social sciences and highlight the main issues that might arise for CESSDA strategy.

5.2 METHODOLOGY

In order to explore the state of play for health data we made literature review to define health data within national contexts or according to specific examples. SND (Sweden) and CNRS-PROGEDO (France) conducted interviews with some key people in public organisations attempting to approach the landscape of health data in both countries. These interviews also contribute to understand how each country collects, uses and stores health data and how the organisations involved deal with them. Various written sources have been utilised.

5.3 SCOPE OF DATA DOMAIN: MAIN UNDERSTANDINGS & NEW ISSUES ARISING

New types of data and also the new opportunities and issues for a research in a digitalised world are discussed in the report "New Data for Understanding the Human Condition - International Perspectives" from OECD Global Science Forum Report on Data and Research Infrastructure for the Social Sciences (OECD, 2013). The report stated that data-driven and evidence-based research is fundamental to understanding and responding effectively and efficiently to global challenges related to the health and wellbeing of populations around the world. Different types

of data, while not new have become newly accessible in the form of electronic records. The category "government and other registration records", including health system registers such as personal medical records and hospital records belong to health data domain. Thus, challenges are following:

- Ensure that there is co-ordination of efforts being made in different parts of the world to develop access to all forms of research data and to capture the potential gains from research use of new forms of data.
- Health data as well as social science data often derives from living persons, this leads to raising legal, ethical, confidentiality and privacy questions that can impede international research.

5.3.1 BOTTOM-UP DEFINITION OF THE DOMAIN

The World Health Organization (1946) defines health as a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity. In that perspective, psychological data for example, could be also considered as part of health data.

The term health data is not precisely defined, although a broad definition has reached a consensus in Europe. In a working paper, written in 2007 by the G29 (or "Article 29 Data Protection Working Party"), it is suggested to maintain a broad approach when it comes to health data. They define them as data that have "a clear and close link with the description of the health status of a person" including alcohol and drugs consumption, etc. The European draft law on data protection widens this definition, stating that health data category includes "every information relative to the physical or mental health of a person, or to the medical service delivered to that person".

CESSDA Archives' approach of health data.

In Table 9, we can see if and how CESSDA service providers define health data in their websites. As shown in the table, most of the archives do not provide a clear definition of health data. This could be due to the fact that many of the data archives only recently started collecting and disseminating health data. Of course, we have to take into account the language barrier, as much information is not provided in English. Moreover, some SPs do not provide health data gave similar results. SPs that provide a definition of health data, they usually provide also data suitable for health research, including for example variables of both physical and mental health, illnesses as well as lifestyle factors, similar and close to the WHO definition. Some archives also divide health data into different sub-categories.

Table 9: Definition of health data per CESSDA member

	Definition of Health data	Country/SP
No information on website	N/A	Belgium/ Belspo
No information on website	N/A	Czech Republic/ CSDA
No information on website	N/A	Denmark/ Danish Data Archive

http://www.fsd.uta.fi/en/data/background/health.html	Description for data on health: Datasets contain data that can be used in health research and they include variables charting, for instance, eating habits and exercise, physical and mental health, and illnesses	Finland/ FSD
http://www.snds.gouv.fr/SNDS/Qu-est-ce-que-le-SNDS	There is no official definition for health data. However, the *National health data system* (SNDS, *Système National des Données de Santé* in French) uses a general definition. This new organisation enables to chain health insurance data; hospital data; medical causes of death; data on disability; and a complementary sample of data from health insurance agencies.	France/ PROGEDO
No information on website	N/A	Germany/ GESIS
No information on website	N/A	Greece/ So.Da.Net
No information on website	N/A	Lithuania/ LIDA
No information on website	N/A	Netherlands/ DANS
No information on website	N/A	Norway/ NSD
No information on website	N/A	Slovakia/ SASD
No information on website	N/A	Slovenia/ ADP
No information on website	N/A	Sweden/ SND
	N/A	Switzerland/ FORS
https://www.ukdataservice.ac.uk/get-data/themes	Topics as varied as the experience of illness, child development, access to care, lifestyle behaviour, subjective physical and mental well-being, diet and nutrition, immunisation programmes and attitudes towards health service provision. Data on health and health behaviour can cover not only a person's status, behaviour, attitudes and expectations but also the provision of health care, including the mechanics of policy making, government expenditure and service coverage	United Kingdom/ UKDA

5.3.2 TRENDS AND STAKES IN RESEARCH

Research using patient data aims in the longer run to improve human health in many forms, via epidemiological studies, public health surveillance, monitoring of drug safety, improved health service management and evaluation of surgical interventions (Sarah & Weale, 2011). This type of research is often carried out by using databases, so that researchers can access quickly large quantities of data, at low cost, and without any interference with research subjects. However, such research often requires data linkage i.e. matching and combining data from multiple databases. Such data linkage cannot be done with fully anonymised data, as it requires some

form of individual identifier to enable matching. Since health data in most cases cannot be fully anonymised without losing valuable information and the possibility to follow up, this legal aspect must be taken into consideration.

The European Commission proposed a comprehensive reform of data protection rules in EU. According to the new General Data Protection Regulation, all EU Member States have to harmonise national legislation by May 2018. This new regulation will have an impact on research data' legal context in general as health data often includes personal data and will therefore be affected by the new law.

There are different barriers regarding data sharing. Regarding public health data, six categories of barriers to data sharing could be identified (Van Panhuis *et al*, 2014):

Technical barriers: i.e lack of data preservation, data that cannot be located, local language used in collection resulting in language barrier, lack of metadata and standards or technical solutions in form of software not available.

Motivational barriers: Individual and institutional motivations and beliefs that restrain data sharing. Among these barriers are no or limited incentives, questionable reliability of the data provider, or disagreement on data use between the data providers and secondary users.

Economic barriers: Lack of resources in the form of human and technical resources **Political barriers:** Structural barriers embedded in the public health governance system, grounded in political or socio-cultural context. Global and national action is required in order to build consensus.

Legal barriers: Legal issues regarding ownership, copyright and protection of privacy.

Ethical barriers: Normative barriers involving conflicts between moral principles and values.

As an example, related to the aforementioned barriers are biobanks and genetic databases that handle personal data in systematic ways and try to deal with legal, ethical and technical systems for sharing personal data.

The importance of access to confidential data is a strategic issue for the research communities in many disciplines e.g. social science, economics and epidemiology. The technological development has made possible new ways of processing very large data files. Along with the evolution of statistical tools for modelling and the possibilities for enriching data by matching different sources, has both contributed to the increasing demand for this type of data. Such highly detailed data are crucial for research, contributing also to public policies evaluation in many crucial domains. Recent developments have for instance underlined the importance of use of highly sensitive medico-administrative data for public health policies. The health sector has for long time been using confidential microdata, mostly from epidemiological cohorts. A basic method for ensuring privacy protection within a specific legal framework for these highly sensitive data has been the extraction of variables. The increasing needs for use and linkage with the medico-administrative databases, as well as the difficulties faced for co-operating

across borders, particularly for the large epidemiological cohorts conducted in many countries, is raising an interest for secure remote access within the research communities[18].

Researchers' needs

In order to investigate researchers' needs within health data, five interviews with health data-experts were conducted, two in Sweden and three in France. In Sweden, one interview was conducted with a university professor expert in the area of public health and one with a first-year post-doctoral researcher in nutrition. The aim was to capture both junior and senior researchers' experience, knowledge and needs. In France, the key informants work in different organisations and are involved in diverse data health domains. They are:

- The Deputy Director of the Institute of research and Documentation in health economics (IRDES, an independent health economic research institute funded mainly by public grants). She contributes to set up national and international surveys in inequalities in health and health care use.
- A Research Director of the *French Institute for Demographic Studies* or INED[19] (a public scientific and technological institute specialised in population studies which produces research at national and international level with the academic and research communities). She works in the research fields of health determinants, the disablement process (measures, determinants, international comparisons), and the connection between gender and health.
- A Professor of Economics at University of Paris-Dauphine (Laboratory of Economy and Laboratory of Economics and Management of Healthcare Organisations).

Concerning researchers' needs the following findings have been identified:
Need for rigorous and accurate health surveys at local and international level; longitudinal studies; matching administrative data with research data; consistent documentation about the data collection; protection of sensitive data and at the same time secure access to these data;

Areas of interest in the domain of health data

Health data field are broad and could be categorised in different ways. The most usual is the randomisation of clinical trials, epidemiological cohort studies or survey studies. Investigate health prevention and behaviour is another way. Some diseases are often used to categorise the subject of research. Administrative data are used by some national health systems like the French one.

Participation in Scientific network for data dissemination

Researchers participate in national and international research networks. They also participate in international surveys like ESS (European Social Survey) and SHARE (Survey of Health, Ageing and Retirement in Europe).

[18] Horizon 2020, Call h2020 INFRAIA 2016-2017 (Integrating and opening research infrastructures of European interest) Data without Boundaries 2, DwB 2 Call INFRAIA-01-2016-2017: Integrating Activities for Advanced Communities- Access to European Social Science Data Archives and Official Statistics
[19] https://www.ined.fr/en/

Dissemination of health data
Researchers in the field of medicine and health are often positive towards sharing data. However, researchers are often concerned with regard to data misinterpretation, as well as with legal aspects of sensitive data. Researchers and research teams have also to deal with authorship issues when disseminate data that have been produced in the context of broader research teams.

Main databases for health data
In Sweden, most databases for health data are provided through Statistics Sweden. It seems that most researchers wish to collaborate with other researchers and make use of existing data. However, in health research this often is associated with personal networking and collaboration between research groups. In France, the device Public Health Database (BDSP for Banque de données en santé publique) consists of libraries, documentation centres, data producers and dissemination agents, which is available to public health specialists. These agencies collaborate in order to develop, supply and distribute data information services in the field of public health. To date, there are forty members participating to BDSP (see the list in Appendix A.3).

Obstacles for researchers' access to health data
Financial restrictions are one of the major obstacles in researchers' access and willingness to share research data. Another obstacle is the development of metadata descriptions and the existing variations between classification standards. An example of this situation is the maternity health registry. This registry was too expensive and difficult to gain access to. The fact that this registry also contains sensitive information, resulted in problems with the original ethical review, which in this case was not approved. Researchers face some obstacles in developing comparative analyses at the international level in view of the fact that each country collects data in a different way.

Sectors of health data of great interest for researchers today and in the near future
Longitudinal data and panel data seem to present currently increasing interest among researchers. The development of validation, translation and sync tools in order to elaborate data of different studies is also considered as important. A key informant argued that in her area of research, epidemiology, environment and climate data present great interest among researchers. Matching health data with administrative records and public health system will be another area of interest in the near future.

5.4 ISSUES ARISING FROM NEW DATA SOURCES AND OTHER ACTORS

The CESSDA service providers offer accessibility, preservation, and re-use of data and related materials. In most cases, individual researchers are the primary data producers and providers, while in some cases they elaborate data from national and international surveys, epidemiological cohorts or administrative data. The situation for health data varies in the data archives in Europe. Some of the archives have a long tradition in incorporating health data in their collections. For other archives, health data is a new domain, where strategies for health data acquisition are under development. The tradition of data sharing also differs among health researchers from country to country.

Legal aspects of health data

One of the key issues with health data is that they often contain personal information. This information has to be handled in accordance with the national laws. This often means that the researcher must anonymise datasets. The EU General Data Protection Regulation (EU, 2016) provides an updated legal framework. This regulation will be incorporated to national legislation by taking into account the need to strengthen individuals' trust and confidence in the digital environment and to enhance legal certainty.

Within the Research Data Alliance, RDA, a Health Data Interest Group has been initiated[20]. The interest group will provide its members with a forum to discuss and highlight the legal, technological, ethical and societal challenges to the adoption of advanced data management and analysis techniques in Healthcare, to exchange opinions and compare experiences. This group will also focus on privacy and security in health data. Amongst other, the group aims at sharing best practices on pseudonymisation, anonymization, differential privacy, and dedicated block chain applications, as well as at developing models for dynamic consent that protect patients while enabling research. CESSDA is also involved in that debate with recent initiatives. The WP6 in SERISS project explores legal and ethical issues of new types of data in detail. CESSDA members involved follow the work in BBMRI ERIC on the development on a GDPR code of conduct for bio-medical data and has initiated a GDPR code of conduct for social science data within the SERISS framework. CESSDA is also setting up a standing committee on legal and ethical issues.

5.4.2 Proportion of data currently archived by the existing data services

In Denmark, **DDA** provides health data and actively promoting the qualification and efficient use of health science data by collaborations with health researchers. Currently DDA cover over 1200 studies and datasets in the health domain in a total of about 10000 datasets according to our estimation.

In Finland, **FSD** provides health data. FSD has 200 datasets containing information of participants' health. Of those, the main discipline is health and/or medical sciences in 132 datasets. Additionally, 12 datasets are classified to psychology and 8 to health policy. The data collected originate from before year 2010 in 148 datasets and 2010 or later in 50 datasets. In total FSD offers about 1300 datasets according to FSD researchers' answers in the survey.

In Sweden**, SND** describes and archive studies and datasets in health and medicine, currently only a small proportion of these studies are archived and disseminated via SND due to the Personal Data Act. Currently, around 150 studies in the health data field are described. A majority of the data files are population based longitudinal health surveys from Sweden. In total SND describes over 1200 studies, of which 552 are in the field of social science and 513 in the humanities.

In Greece, **SoDaNet** does not provide health data.
In Czech Republic, **CSDA** does not provide health data.

[20] https://www.rd-alliance.org/group/health-data/case-statement/health-data.html

In Netherlands, **DANS** provides 402 datasets in the fields of Life sciences, medicine and health care.

In Germany, **Gesis** provides various health datasets under the category of Medicine and under the topics of Health, General Health, Health Policy, Health Care and Medical Treatment, specific diseases and medical conditions etc. The European System of Social Indicators, covering the EU-27 member states, Norway and Switzerland as well as Japan and the United States, also offers rich data in the health domain.

In France, **ADISP** (National Archive of Data from Official Statistics) works within **PROGEDO (CNRS).** ADISP aims to disseminate surveys, studies and databases produced by INSEE (the French National Institute of Statistics and Economic Studies), and other public institutions such as the French National Institute of Health and Medical Research (Inserm), the Institute for Research and Information in Health Economy (IRDES) and the Ministry of Public Health. Its catalogue hosts more than 100 national surveys and reports in open access for the whole scientific community. This data is archived in the *Quetelet-PROGEDO diffusion* database and by each public institution that developed the survey. In addition, the French health insurance offers a complete and detailed database on the data of patients and the organisation of the healthcare system. The French health insurance provides three sets of health data: 15 thematic databases referred to a special purpose; a General sample of beneficiaries (EGB in French) of the population protected by French health insurance, based on a survey at the one percent on the social security number of French health insurance beneficiaries (around 660,000 beneficiaries); and a single database of beneficiaries on consumption of care. This individual data of beneficiaries is available on 3 years beyond the current year.

In the UK, the **UK Data service** holds 1332 datasets classified under the health theme.

In Lithuania, **LiDA** provides 22 health datasets in a total number of about 300 datasets. Health datasets concern drug abuse, alcohol and smoking, health services and medical care, nutrition, diseases etc.

In Slovenia, **ADP** provides 30 health datasets in a total number of about 600 datasets. Health datasets concern indicatively drug abuse, alcohol and smoking, health care and medical treatment, public opinion surveys about health and health services etc.

In Switzerland, **FORS** provides about 60 datasets in the topic of health out of a total number of about 10000. Indicatively, these datasets concern with diseases and medical conditions, accidents and injuries, abortions, drug abuse, alcohol and smoking, health care and medical treatment etc.

5.4.3 TYPE OF AGREEMENTS WITH PRODUCERS THAT IMPACT THE COVERAGE

Individual researchers or institutional data producers such as INSEE (the French National Institute of Statistics and Economic Studies) or DREES (a branch of the central administration of the social ministries) in France are the main providers of health data to the data archives. However, there are also agreements with producers that impact coverage of health data in the

CESSDA service providers. Some SPs have made agreements with data producers in order to provide them systematically with data, including in some cases health data.

Swedish National Data service (**SND)** has an agreement with Swedish Cohort Consortium in Sweden to describe study information and metadata for the projects involved. However, due to personal identifiers in the data, SND just describes the data. SND also is a certified Trusted Digital Repository and listed as a recommended repository by PLOS Journals.

The Danish data archive **(DDA)** has created an electronic form with the aim that all research projects conducted in Denmark should fill in some basic information and thereafter be assessed with regard to reproducibility of the data. If the data are assessed as suitable for reuse, DDA makes an agreement about deposition with the researcher or institution in question. In some cases, DDA offers external help with documentation of research material that are of particular interests for the research community. DDA is deeply involved in this work, collaborates with the health research environment, and has established a network of researchers in health science.
The Norwegian Centre for Research Data (**NSD)** collaborates with the Statistics Norway (SSB) in order to facilitate and distributing SSB`s data to Norwegian research institutions.

In the UK, the **UK Data service** collaborates with data producers and data owners of some important health data sources named as Key data. The Health and Social Care Information Centre and the Centre for Longitudinal Studies regularly deposit data at UKDA are two important producers of health data that regularly deposit data.

In France, **ADISP** (National Archive of Data from Official Statistics) signed agreements with INSEE (the French National Institute of Statistics and Economic Studies), statistical departments of ministries and other public institutions to maintain and expand its catalogue. Since 2017, the French National System of Health Data **(SNDS)** enables to chain different kind of data (as we can see below: health insurance data; hospitals data; etc.). The SNDS's main purpose is to provide these data in order to promote studies, research or evaluation of a public nature. **IRDES** (Institute of research and Documentation in health economics) has signed agreements with the French health insurance, the Direction of Research, Studies, Evaluation and Statistics (DREES[21]); and the Ministry of Health to develop national surveys.

In appendix 1 a list of actors impacting coverage outside CESSDA in the health domain, is provided.

5.4.4 National level policies & related strategies

The development towards open access to research data is increasing at the National French system. The law 26-01-2016 established the modernization of the French health system, by the creation of the National System of Health Data **(SNDS)**. One of a kind in Europe, the SNDS is a major step forward to analyse and improve the health of the population. SNDS disposes of

[21] The Direction of research, studies, evaluation and statistics (Drees) is a direction of the central administration of health and social ministries. It is under the supervision of the Ministry of Economy and Finance, the Ministry of Solidarity and Health, and the Ministry of Labour.

various types of health data, such as health insurance data, hospitals data, medical causes of death, data on disability (from 2018 onwards), a complementary sample of data from health insurance agencies (from 2019 onwards). Therefore, the SNDS contributes to develop the health information, as well as to implement health policies by the *National Institute of Health Data* (INDS). This institute runs the access to the health data for public organisations and researchers. Access by companies and health insurers is strongly framed, since the SNDS mainly holds personal health data. Processing of such data should be strictly framed to protect the privacy of individuals.

The French health insurance has an exhaustive record of the health care system because all the population (around 67 million) has a social security number, which is used in each medical intervention, even in the private health services. However, this data needs to be processed to develop research in an accurate way. In this context, many public organisations produce surveys gathering health data through sampling, for instance:

- the ESPS Survey in Health, Health Care and Insurance Survey (biannual survey, last wave 2006), runs by the IRDES;
- two surveys on health and disability in 2008-2009 and one on health and old ages dependency in 2016, run by the DREES
- the National Health Survey (or ten-year health survey), runs by the INSEE until 2003.

In the case of the National survey runs by INSEE, a sample of the population is interviewed at home to better understand other variables beyond the health care, such as social environment or family. The national statistics of public health are built by this method, which is also used in some international surveys. Actually, based on national health interview surveys, EUROSTAT proposes to develop comparable modules of questions to allow European comparisons. Subsequently, this data is matched with the administrative records (health insurance data, hospital records, etc.), in which the identity of patients is anonymised. Some experts wonder if the French health data could become a sort of "Big data" as the health insurance has exhaustive records, which are open for the research community. However, other experts argue that this data has not yet been processed and for that, researchers should conduct high-quality research to treat Big data.

In Sweden, the development towards open access to research data took a significant leap forward in 2015. The Swedish Research Council´s (2015) provided a set of national guidelines for open access, including research data. A couple of recommendations are of particular importance regarding access to research data, such as the establishment of a central co-ordinating agency at national level. Finland and Denmark provide examples with similar structure. The Danish Data Archive, a CESSDA SP, has been merged with the Danish National Archives and disseminates digital data arisen from research domain and public administration, including health data. Moreover, the Swedish Research Council along with various research organizations (2015) provided with a report regarding the Swedish Government's Research and Innovation Bill. This report stresses the importance of an e-infrastructure to support the entire research process, making accessible data arisen from research domain and public administration, including health data. This e-infrastructure will have an impact in the health data domain, increasing the pressure to ensure privacy protection of sensitive data.

6. OFFICIAL STATISTICS

6.1 GENERAL OVERVIEW OF THE OBJECTIVES AND ORGANISATION OF WORK

6.1.1 GENERAL UNDERSTANDINGS OF "OFFICIAL STATISTICS"

While the overall meaning of "official statistics" (OS) is rather straight-forward at a global and European level, the definition varies across countries. As underlined in the outputs of the DwB project:

"in the United Kingdom, official statistics is understood to be statistics produced by the Office for National Statistics (ONS) plus the statistics produced by any other organisations involved in providing a public service; in France, where the word "public" is used instead of "official", the definition includes all data productions originating from statistical surveys of Institut National de la Statistique et des Études Économiques (INSEE), and the use of data collected by all organisations with a public service mission".

Considering the expression can concern a broad range of types of data and agencies, discussing general understandings of the expression OS here is essential. On an international level, the UN Statistics Office (2014) suggests guiding principles, hence providing insight on the purpose and range of data involved. It states that OS, based on transparency, trust and professional considerations, are "*an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation*".

These understandings of OS fit the Administrative Data Research Network views. This professional organisation suggests OS are meant to give in-depth and accurate pictures of society. They also emphasise the fact such information is extremely valuable for social and economic research since scientific findings based on these tools have the potential to advise future government policy as well as influence how politicians and others evaluate existing policies[22].

Whether referring to understandings of OS by the OECD or the Administrative Data Research Network, data may be drawn from all types of sources, whether statistical surveys or administrative records. Along the lines of this understanding, with the goal of widening CESSDA's data perimeter by furthering co-operation between "data archive services" (DAS) and OS, here we will mainly focus on micro-data, as opposed to other types of data, namely aggregated or macro-data. One may refer to definitions of the OECD[23] to define microdata, in which case:

"Microdata is the file consisting of the set of records where each record represents individual statistical unit. The term microdata can refer to data about an individual person, household, business or other

[22] https://adrn.ac.uk/admin-data/admindata/

[23] Microdata-access-final-report-OECD-2014.pdf: Microdata (or statistical microdata) Source: UNECE, Managing Statistical Confidentiality & Microdata Access - Principles and Guidelines of Good Practice, 2007. p. 1, http://www.unece.org/fileadmin/DAM/stats/publications/Managing.statistical.confidentiality.and.microdata.access.pdf

entity. It may be data directly collected by the NSO or obtained from other sources, such as administrative sources."

This optic falls in line with the goals of the CESSDA Official mission statement. The latter specifies that the consortium aims at providing a comprehensive, distributed and integrated social science data research infrastructure, facilitating access to social science data resources for researchers, regardless of the location of either researcher or data. Thus statistical micro data are by definition part of the scope which makes a case for continues and improving co-operation etc.

At the European level, purpose and range of data line up with one of the core infrastructures of OS, that is represented in Europe by ESS, the European Statistical System, including Eurostat[24]. ESS works in close relationship with National Statistical Institutes (NSIs) and other national authorities responsible in each Member State for the development, production and dissemination of statistics[25].

However, when considering OS on a national level, the scope and the number of actors can be more or less broad, as we mentioned before[26]. While contributions of numerous, sometimes small, agencies make the systematic tracking and centralization of data produced difficult, the uneven landscape especially blurs definitions of what is considered as OS. The variations in the meaning of the expression OS makes it also somewhat unclear what type of data can be accessed nationally. As noted in the Data without boundaries project (DwB), "*differences in the organisation of national statistical systems, related to their degree of functional and/or geographical centralization, potentially affect access to national data centres even more so*"[27]. With that being said, going off of previous findings, it seems safe to consider that OS revolves around statistics produced by governments and boards, private corporations, watchdogs or regulators providing some kind of public service.

In the context of this report, data from OSs refers to:

- censuses, i.e. measurements of people and households in a given country
- surveys, i.e. collection of information about citizens
- business and economic data, i.e. collections of information about members of a population of companies, obtained through surveys and/or administrative registers (Tubaro, Cros & Silberman 2012)
- administrative data, i.e. data collections built by government services and agencies when registering people or carrying out transactions, or for record keeping – usually when delivering a service (e.g. social security payment records, educational attainment records, health records, court records, tax records[28]).

[24] http://ec.europa.eu/eurostat
[25] http://ec.europa.eu/eurostat/web/european-statistical-system/overview
[26] DwB D3.3 Researcher accreditation - current practice, essential features, and a future standard p. 4.
[27] *Ibid.*
[28] https://adrn.ac.uk/admin-data/admindata/

6.1.2 DESCRIPTION OF THE OBJECTIVES

As implied in the previous lines, statistics are now produced by a wide range of agencies. This evolution complies with societal demands as OSs play an expert role. If the downside of such a massive production of data is recent disillusions of social actors, new formats and data production trends shift the focus of what statistical data may become tomorrow. The increasing use of web data or transactional data produced by social media and private companies, as well as the development of administrative data use are important to consider, since such evolutions raise unprecedented issues in terms of preservation, documentation and access for the social sciences, as well as needs for researchers.

These changes bring up issues such as:

- What are the legal frames for new sources of data and new resources for researchers and DAs?
- What type of co-operation with researchers could be foreseen? How to share the value-added products that researchers can contribute, and that are partially derived from official micro data? How can other researchers access the data to validate published results?
- More and more official micro-data is more accessible (with various points of access and fragmentation of the field) and potentially more used, but overall more disperse. How does this affect the field?
- When co-operation is efficient, why has it succeeded?

There are some recurrent topics that were partially addressed in past projects such as DwB but deserve more attentions when trying to map the whole landscape of co-operation and address the uneven situation in different European countries regarding access to official microdata.

Some of those topics will be addressed to see if there is potential to move forward in broader range of countries, based on examples and existing good practices, e.g.:

- Questions about existence, persistence of microdata, and access options in various countries, that is, enabled with the descriptive metadata?
- Types and ranges of data that is accessible under different access regimes?
- Value added microdata products, such as cumulative files of continuous surveys?
- Access to linked and continuous administrative microdata?
- Value added services, such as mapping of official classifications, users' conferences, literature related to data, etc.?

Therefore, the goal of the sub-task will be to highlight the recent evolutions of the statistical field, in regards with the scope of the domain, researchers' needs and emerging datasets types. It should extend the investigation to administrative data and big data, as well as co-operation between archives and NSIs to the Nordic countries, poorly enquired in past reports, especially when such data formats are centralised by the NSI. Establishing such changes should help identify new trends and upcoming developments of the field.

6.1.3 METHODOLOGY

The findings are initially based on the extensive research that has been conducted in the past few years. We carefully review what has been brought forward by previous reports, notably in the cases of the CESSDA PPP and DwB projects. Namely, work conducted during the CESSDA PPP[29] and the DwB[30] project allowed to obtain a state-of-play of existing relationships between archives and the National Statistical Institutes, types of co-operation and in particular whether the archive catalogue includes official statistics or not. We will include other sources of information, may it be institutional reports or scientific literature, reporting recent evolutions of the field, highlighting researchers' need or provide a general picture of the national or individual landscape of statistical micro-datasets.

To search the field of new and emerging archiving services, and spot possible data providers, a web investigation is conducted by navigating online and identifying data centres, within the ERA countries - i.e. scope of countries potentially of interest for CESSDA - fostering official statistics of some kind. Moreover, interviews with experts in the field of OS were conducted by ADP in order to explore researchers' needs.

6.2 SCOPE OF THE DATA DOMAIN: MAIN UNDERSTANDINGS & NEW ISSUES ARISING

Comparing data deposited in various archives, may it be within a country or cross-nationally, can be difficult due to the range of terminologies employed and the different classifications used for similar domains. From a research perspective, this lack of homogeneity complicates comparability. From a user perspective, this situation leads to confusions concerning the type of data archived. To clarify terminologies used amongst Member country archives, task 3.4 tackles this problem by investigating online the manners in which formats of data are defined by archiving services.

Once these elements are presented, this section explores the main understandings of the domain as well as the new issues arising. For CESSDA to prepare for the future, it is necessary to consider the challenges that lie ahead and establish the main difficulties that may arise. Likewise, it is important to consider how researchers' needs are evolving.

The domain of OS is different from the other data domains under study in task 3.4. Work has been previously conducted by the OECD Expert Group for International Collaboration on Microdata Access[31]. Likewise, previous European projects on microdata carried out extensive

[29] Work package 10 of the CESSDA PPP was engaged in work on access mechanisms and availability of Official Statistics across the European Research Area D10.1:
http://www.cessda.org/project/doc/D10.1_Audit_of_access_mechanisms_and_official_statistics.pdf

[30] Data without Boundaries - DwB project had a mission to support equal and easy access to the rich resources of official microdata for the European Research Area, within a structured framework where responsibilities and liability would be equally shared. It was implemented by a big Consortium of 29 partners, NSIs, Data Archives and universities www.dwbproject.org

[31] Between 2005 and 2007, the OECD conducted exploratory work to investigate the feasibility of making official microdata more accessible to policy makers and analysts (cf. "Study on the Feasibility of Micro-Data Access for the OECD" (STD/CSTAT(2007)3/ANN). The OECD built on this earlier effort in the following years. It prospected more general issues to facilitate microdata access.
https://www.oecd.org/std/microdata-access-executive-summary-OECD-2014.pdf

work. In DwB in particular, terminology was decided upon, laying the floor for common grounds of understanding, furthering the ultimate goal of that project, i.e. improve co-operation between archives and NSIs. But while a playing field was established to define OS in both cases, the terminology adopted is the by-product of top-down logics. DwB based their reporting on the wording used in the European regulation of access to confidential data for scientific purposes (European Commission 2013a) and the OECD's Expert Group for International Collaboration on Micro-data Access (2014). These two references perpetuate bottom-down definitions. To avoid disregarding past efforts, these proposals will be considered below against the finding of the online investigation carried during SaW meant to see how data centres approached the datasets in their holding.

While the investigation of terminologies given by archives is a relevant goal, difficulties to establish such a proposal rapidly appeared during the web investigation: either websites aren't accessible in English, or, more often than the latter, they do not contain this information. The information contained on the CESSDA partners' websites eventually state the purpose and/or the mission of the institution regarding OS, but they do not supply an actual definition of the field. NSI's were however more likely to offer a definition of the data they were handling and that they considered like OS. The table below synthesises both NSI's and CESSDA's partners understandings when such information is provided.

Table 10: National Statistical Office (NSO) or National Statistical Institute (NSI)

Definitions	Sources
Official statistics refer to public information which is produced for the benefit of the society and is funded by the state budget under the official or European Union statistical programme. Official statistics are equally accessible to everyone and enable the consumers to make the necessary decisions in their private or business lives. Official statistics comply with international classifications and methodologies and meet the principles of impartiality, reliability, relevance, cost-effectiveness, confidentiality and clarity. European Union official statistics are regulated by the quality criteria established in the European Statistics Code of Practice. There are two producers of official statistics in Estonia – Statistics Estonia and Eesti Pank (central bank of Estonia).	*Statistics Estonia and Eesti Pank (central bank of Estonia)*[32]
Statistics in the information society "Information" has become an important economic and social factor. An effective "information culture" is as vital to the success of any institution as the procurement and selection of relevant information is to the success of any manager. The huge information requirements of our society have led to the development of individual branches for the supply of information and for the development of entirely new media enabling a rapid, international transfer of information and convenient information processing. The information explosion that has taken place as well as the increased demands made on management have made it necessary to filter out the relevant information from reliable sources from an enormous supply. It is the role of Statistics Austria to provide reliably collected and expertly analysed political, social and economic information. While statistics were originally created for administrative purposes and to form the basis of political decision-making, their application and use for this purpose has become of increasing importance for the general public. By offering tailor-made services, Statistics Austria tries to meet individual requirements and to provide the requested information to users in an easily accessible format and as quickly as possible. With EU entry, another function of Statistics Austria has gained in importance from the users' point of view: its function to mediate between Austria and EU data as well as to co-ordinate the pan-European harmonisation process.	*Statistics Austria*[33]

[32] www.stat.ee/what-are-official-statistics-and-how-are-they-produced
[33] http://www.statistik.at/web_en/statistics/index.html

Official Statistics of Finland (OSF) are a comprehensive collection of statistics describing the development and state of society. They comprise nearly 300 sets of statistics on 26 different topics. The basic data of the Official Statistics of Finland are available to all users free of charge. [34]	*Official Statistics of Finland (OSF)*
The term "official statistics" includes all material generated by statistical surveys, as specified in the list determined every year in a ruling by the Ministry for the Economy, and the use of data collected by government administrations, public or private bodies with a public service role for purposes of general information. *The design, production and dissemination of official statistics are conducted with full professional independence by the official statistical system, and by producers approved by the National Council for Statistical Information (CNIS) or the Official Statistics Authority (ASP).*[35]	*National Institute of Statistics and Economic Studies (INSEE)*
The Croatian Bureau of Statistics, as a principal producer of official statistics, continuously monitors and applies world and European statistical standards, particularly with regard to the statistical classifications, which are one of the basic tools for the production of statistical data.	*Croatian Bureau of Statistics* *Croatia*

Table 11: Definition of statistical data per CESSDA member (where information was available)

	Definition of data/ official stats/ official data	**Country/SP**
STI rather than OS?	The Federal Science Policy's scientific and technical information department provides all the information on research statistics in Belgium. (..) STI is an acronym for 'science, technology and innovation'. This title covers a very wide range of activities. Within this conceptual framework, 'Research and Experimental Development (R&D)', which is present in all economic sectors, is of major importance.	Belgium/ Belspo
Translated from French (no English version on the website) Purpose of OS rather than definiton	The analyses and key figures enable you to get an idea of the economy of Belgium. We must of course consider them in a wider European and global economic context. Many data on economic activity comes from international organisations such as the OECD (link is external), Eurostat (link is external), the IMF (link is external) ... (...). The main mission of the Directorate General Statistics of the FPS Economy (link is external) (Statistics Belgium), is to collect, process and disseminate statistics on the Belgian company. Many statistics are interesting and may even be important when you want to start a business. You will find these numbers on the website of the Directorate General of Statistics, in several different categories (each with its sub-sections): population; the labor market and living conditions; economy; traffic and transportation; environment; energy (On Belgium.be)	
For further information on purpose archiving system: http://www.fsd.uta.fi/lehti/en/12/tilastokeskus.html	The official statistics of Finland are divided into 28 subject fields, which cover the phenomena included in the UN international statistics series: population, social, economic and other statistics. In addition to Statistics Finland, about 20 other ministries and state institutions produce national statistics.	Finland/ FSD
Description of service and data types rather than a definition of the	The German Microdata Lab (GML) collects microdata of official statistics. It offers research based services concerning the data and develops instruments for the implementation of social scientific	Germany/ GESIS

[34] http://tilastokeskus.fi/meta/svt/index_en.html
[35] http://www.insee.fr/en/insee-statistique-publique/default.asp?page=statistique-publique/statistique-publique.htm

domain	concepts. The German Microdata Lab (GML) collects microdata of official statistics. It offers research based services concerning the data and develops instruments for the implementation of social scientific concepts. Service for microdata from official statistics: Metadata for Scientific Use Files in MISSY (Microdata Information System): (German Microcensus, EU-SILC (EU Statistics on Income and Living Conditions); EU-LFS (EU Labour Force Survey); AES (Adult Education Survey); CIS (Community Innovation Survey); SES (Structure of Earnings Survey) + European Microdata+ Microcensus Trendfile (cumulation of microcensus data 1962-2006) + Microdata Tools + Information on further microdata (Income & Expenditure Survey; Population and Occupation Census 1970 (VZ 1970); GDR-Data; Workplace and Occupation Census (AZBZ); Information on further Official Microdata)	
Mission of ADP & use of OS rather than definition of the domain	Official statistics data are a valuable source for sociological research. Users of data have become more aware of that and their interest in accessing official statistics microdata has increased. Therefore, Social Science Data Archives (ADP) have provided an additional support for their work, in addition to the distribution of the Statistical Office of the Republic of Slovenia (SORS) metadata and microdata. The distribution of Labour Force Survey, Time Use Survey, Crime Victim Survey and Household Budget Survey data has been going on since the establishment of ADP.	Slovenia/ ADP

The few definitions retrieved fall in line with the principles of the UN Official Statistics Office, in the sense that data produced is public trustful, professionally produced information, meant to serve a greater good, bringing together a wide range of data types, whether statistical surveys or administrative records. This prolongs recent discussions: "*Besides statistical purposes, the potential of microdata for policy and scientific purposes has been increasingly recognised over recent years. Their analysis being facilitated by technological developments, microdata are extremely valuable as they provide the possibility to assess the underlying structure and causal links of the studied phenomena. At the same time, the calls for governments' transparency and accountability are influenced not least by the open…*" (Eurostat 2016).

6.2.1 NEW DATA FORMATS: ADMINISTRATIVE, BUSINESS AND ECONOMIC DATA

OECD Social Science GSF Expert Group evokes new formats of data that can be associated to administrative data and business and economic data[36]. In the field of administrative data, some new formats of data, as listed by the OECD Expert group, are presented in Table 12 below.

Table 12: New formats of data in administrative data

Examples	Detailed categories	Broad category of data
Income tax; tax credits	Individual tax records	**Government transactions**
Corporation tax; sales; tax; value added tax	Corporate tax records	
Tax on sales of property; tax on value of property	Property tax records	

[36] http://www.wisc.warwick.ac.uk/files/2914/4613/9650/WISC_Peter_Elias_11_February_2015.pdf
http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf

State pension; hardship payments; unemployment benefits; child benefits	Social security payments	
Border control records; import/ export licensing records	Import/ Export records	
Registers of ownership	Housing and land use registers	**Government and other registration records**
School inspection; pupil results	Educational registers	
Police records; court records	Criminal justice	
Registers of eligible persons	Social security registers	
Voter registration records	Electoral registers	
Employment census records: registers of persons joining/leaving employment	Employment registers	
Births; marriages; civil unions; deaths; immigration/emigration records; census records	Population registers	
Personal medical records; hospital records	Health system registers	
Driver licence registers; vehicle licence registers	Vehicle driver registers	
Political parties; charities; Clubs	Membership registers	

Commercial transactions are another new type of data. Interactions with business are leaving additional digital prints that can be either associated to the realm of transactional data (and therefore big data), or the realm of business and economic data. OECD Social Science GSF Expert Group listed three categories in particular:

1. Store cards (supermarket loyalty cards; membership cards)
2. Customer accounts (utilities; financial institutions; mobile phone uses)
3. Customer records (product purchases; service agreements)

Some of the examples hereinabove, i.e. mobile phone uses, are more often exploited by researchers in the field of social media or in relationship to Internet uses and therefore appear closer to big data. What this latter example mainly demonstrates though is that data fields are increasingly intermixed.

6.2.2. CHALLENGES BROUGHT FORWARD BY BIG DATA

- **The big data turn in Official Statistics**

Big Data is one of the key assets of the future. Mastering the creation of value from Big Data will enhance European competitiveness, will result in economic growth and jobs, and will deliver societal benefit. Strategic investments by industry, the public sector and governments, accompanied by forward-looking policies, will enable Europe to take the lead in the global data

economy and to reap immense societal benefits from the unique opportunities offered by Big Data Value.[37]

Big Data has many expected benefits: its potential to spur innovation, deliver better services for less money, improve planning, increase transparency, reduce corruption, and reveal patterns and insights. Deriving value from Big Data is a critical factor for social innovation, for the provision of meaningful public and corporate services and for the optimization of decision-making processes at various levels and in manifold contexts. Thus, the availability of high quality data assets and technologies that are required for acquiring, managing, and exploiting Big Data is of major importance for entities involved in data value chains as well as the wider social environment on which they operate.

There is no doubt that Big Data is revolutionising business today, but there is still no unanimous definition of it. There is a number of definitions offered by the world's biggest and most influential high-tech organisations[38]:

- Gartner and Gartner in 2001 report: the increasing size of data, the increasing rate at which it is produced and the increasing range of formats and representations employed. This report predated the term "Big Data" but proposed a three-fold definition encompassing the "three Vs": Volume, Velocity and Variety. This idea has since become popular and sometimes includes a fourth V: veracity, to cover questions of trust and uncertainty.

- Oracle: Big Data is the derivation of value from traditional relational database-driven business decision making, augmented with new sources of unstructured data.

- Intel: Big Data opportunities emerge in organisations generating a median of 300 terabytes of data a week. The most common forms of data analyzed in this way are business transactions stored in relational databases, followed by documents, e-mail, sensor data, blogs, and social media.

- Microsoft: Big Data is the term increasingly used to describe the process of applying serious computing power—the latest in machine learning and artificial intelligence—to seriously massive and often highly complex sets of information.

- The Method for an Integrated Knowledge Environment open-source project: The MIKE project argues that Big Data is not a function of the size of a data set but its complexity. Consequently, it is the high degree of permutations and interactions within a data set that defines Big Data.

- The National Institute of Standards and Technology: NIST argues that Big Data is data which "exceed(s) the capacity or capability of current or conventional methods and systems." In other words, the notion of "big" is relative to the current standard of computation.

[37] European Big Data Value CPPP - Strategic Research and Innovation Agenda - April 2014
[38] Undefined by Data: A Survey of Big Data Definitions accessible at: arxiv.org/abs/1309.5821

In 2013, ESS (European Statistical System) took note of a "big data turn" by adopting the Scheveningen Memorandum (DGINS 2013). Three goals were set forward:

- acknowledge that Big Data represents new opportunities and challenges for Official Statistics;
- encourage the European Statistical System and its partners to effectively examine the potential of Big Data sources;
- agree on the importance of following up the implementation of this memorandum by adopting an ESS action plan and roadmap by mid-2014.

By acknowledging the digital transformation and putting big data on the roadmap of wider government national and international strategies, ESS (2014) leaped onto the bandwagon of upcoming challenges and suggested manners in which OS could redefine its role in this new context:

"A digital transformation is taking place across the globe. e ever increasing availability of data is a trend that is of strategic relevance for official statistics. There is a need to assess and interpret the meaning of these data in intelligent and interactive fashion. These new data sources are a huge opportunity to improve the timeliness and relevance of official statistics as well as to lower response burden. On the other hand, there will be more competitive pressure from new data producers which can eventually change the role of official statistics. We have to answer the core question: what is the future role for a reliable and high-quality information infrastructure in such an environment?"

Many national initiatives pursued by the Statistics Netherlands (with traffic loop and social media data, notably), Statistics Ireland, CBS Netherlands, ISTAT Italy, ONS UK, CSO Ireland, Statistics Finland, SURS Slovenia[39] follow in these footsteps by linking big data and statistics.

- **Two specific changes brought forward by big data**

1. The first type of change can be categorised as hypothetical, more along the lines of what could be. In this case, changes concern the context of data production, that is to say the manners in which organisations producing official statistics operate (Struijs, Braaksma & Daas, 2014).

New collaborative opportunities are emerging between different types of data providers, producers and archives, as the development of big data brings together private corporations, watchdogs, National Statistical Institutes and academics. This development questions the role statistical institutes will take on in a near future. Going against what we previously said on regarding "statistical disillusion", Strujis, Braaksma and Daas (2014), support the idea that NSIs will manage to uphold the provision of high-quality and impartial statistical information to society, and precisely make this quality their force. They claim:

"The collaboration between the various stakeholders will involve each partner building on and contributing different strengths. For national statistical offices, traditional strengths include, on the one hand, the ability to collect data and combine data sources with statistical products and, on the other hand, their focus on quality, transparency and sound methodology. In the Big Data era of competing and multiplying data sources, they continue to have a unique knowledge of official

[39] http://www.statistiques.public.lu/fr/agenda/detailagenda/2015/10/SKALIOTISWorldstatsdaySTATEC.pdf

statistical production methods. And their impartiality and respect for privacy as enshrined in law uniquely position them as a trusted third party. Based on this, they may advise on the quality and validity of information of various sources. By thus positioning themselves, they will be able to play their role as key information providers in a changing society."

Another possible opportunity created by the advent of big data is that of improving the accuracy, timeliness, and relevance of economic statistics at a lower cost than expanding existing data collections[40].

2. These somewhat hypothetical changes are tailed by very real dilemmas that need to be faced on different fronts[41]:

- **Legislative**, i.e. with respect to the access and use of data. Web data freely available on the web escapes existing legislation.
- **Privacy**, i.e. managing public trust and acceptance of data re-use and its link to other sources. Users producing transactional data are possibly unaware their data can be further exploited.
- **Financial**, i.e. potential costs of sourcing data vs. benefits. Big must be stocked, bought in some occasions, and there is no legislation to regulate acquisition of external data. The process of exchanging, sharing, integrating and joining Big Data is cumbersome and resource demanding.
- **Management**, e.g. policies and directives about data management and protection. The additional information generated by big data, pouring over into NSI's raises management and protection policies issues, plus possible long-term stability problems.
- **Methodological**, i.e. data quality and suitability of statistical methods. Traditional methods developed for small samples are being trailed. Plus, data are likely to be selective, non-representative, or feeding off the digital divide.
- **Technological**, i.e. issues related to information technology. Dedicated and specialised computing infrastructures are required to cope with Big Data to enable processing and speed up analysis of large amounts of data. Certainly, for the exploratory phase, during which the content and structure of Big Data sets has to be understood, fast technology certainly speeds up this process and more quickly enable the revelation of their use for statistics. In any case, ensuring the interoperability and the transferability of technological solutions applied across different domains is not a trivial process. The evolution of technologies around Big Data is happening at a pace that makes the adoption of new technologies in an established value chain very tedious.

In order to contextualise and analyse the aforementioned challenges, we can identify two distinct aspects of challenges faced by participants in Big Data value chains, or entities that aim to adopt and exploit Big Data technologies. On the one hand, Big Data poses a technology shift burden on all aspects of a data value chain and particularly on the following steps:

- Generating and/or acquiring data
- Storing and curating data collections
- Processing and analysing data

[40] http://unstats.un.org/unsd/trade/events/2014/beijing/Steve%20Landefeld%20%20Uses%20of%20Big%20Data%20for%20official%20statistics.pdf
[41] http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2013/Topic_4_Daas.pdf

- Visualising and using data
- Providing meaningful services over data

On the other hand, there is an orthogonal relation between Big Data enabling technologies and various communities that can benefit from the incorporation of such technological solutions in their business models. These generic technologies enabling the maintenance, usage and exploitation of Big Data are of varying importance and are applied in different ways across different domains and communities. Each application domain bears its own characteristics and requirements and demands different technological and data assets to be used in order to effectively use the underlying information. Furthermore, the aspect of cross-community data asset sharing is a factor of significant importance, and one that cannot be assessed straightforwardly.

6.2.3 CESSDA'S INVOLVEMENT IN BIG DATA INITIATIVES

1. Big Data Europe project

CESSDA is a beneficiary in the "Big Data Europe project - Empowering Communities with Data Technologies" (BDE project a 3-year Horizon 2020 CSA focused on providing an integrated stack of tools to manipulate, publish and use large-scale data resources. Big Data Europe focuses on two clearly defined co-ordination and support measures:

1. Engaging with a diverse range of stakeholder groups representing particularly the Horizon 2020 societal challenges Health, Food & Agriculture, Energy, Transport, Climate, Social Sciences and Security; collecting requirements for the ICT infrastructure needed by data-intensive science practitioners tackling a wide range of societal challenges (co-ordination).
2. Designing, realising and evaluating a Big Data Aggregator platform infrastructure that meets requirements of diverse interest groups (support).

BDE project embraced Gartner and Gartner definition of Big Data which encompassing the "three Vs": Volume, Velocity and Variety, and sometimes including a fourth V: veracity, to cover questions of trust and uncertainty. CESSDA'role is to co-ordinate the SC6 Interest Group, as well as potential users of Big Data in the fields of social sciences and humanities (SSH). Furthermore, CESSDA has been working on the build-up of this interest group, collecting its requirements, assisting the building of an ICT Big Data infrastructure as an access point for SSH, exploring and evaluating the input data, and discovering the implications for the future of Big Data in SSH.

2. SMARTPolicy proposal

CESSDA was engaged in the development of the SMARTPolicy project proposal within the H2020 programme of the EU. The proposal was submitted on 20 February 2017. Data-driven decision making has become an essential component for different practices across various fields of human actions; from educational practices (Mandinach 2012) to environmental issues, and has received substantial attention in terms of policy and financial support. There is also a significant difference between data-driven and theory driven policy development; while for the

data-driven development of indicators that will later be transformed into the policies, availability of data is the crucial indicator itself, for theory-driven approach the

The opportunities associated with data and analysis in different organisations are used to help enterprises better understand their internal processes and market, and subsequently make timely business decisions. Underlying data processing and analytical technologies connected to business intelligence and analytics includes practices and methodologies that can be applied to various high-impact applications such as e-commerce, market intelligence, e- government, healthcare, and security (Chen et al 2012). Over the past decades business intelligence, analytics and the related field of Big Data analytics have become increasingly important in both the academic and the business communities. More recently, Big Data and Big Data analytics have been used to describe the data sets and analytical techniques in applications that are so large (from terabytes to exabytes) and complex (from sensor to social media data) that they require advanced and unique data storage, management, analysis, and visualization technologies. In addition, there is abundance of unstructured, novel forms of data coming from various online social media (forums, online groups, web blogs, social networks and social multimedia sites, etc.) that can offer insight into various issues, i.e. socio-political sentiments or opinions from different stakeholder groups.

Growing importance of evidence based decision-making particularly in the public sector, and emphasis on citizens involvement and more transparent information sharing, has created a need for effective and reliable policies to take possible indicators that are theoretically supported, and data availability is considered as one of the aspects involved in the process (Niemeijer 2002). Another particularity of data-driven decision-making having profound impact on policy development involves the use of data to identify patterns of performance that can reveal strengths and weaknesses focus is on selecting the best connected to established goals, and therefore influence the planning of instructional practices for all participants.

CESSDA and its Linked Third Parties (Service Providers: ADP, EKKE, CSDA, ICS ULisboa, NSD, UKDA) were in charge of development of the WP2 - Methodology for Data-Driven Policy Development. The rationale behind it was based on the fact that data measuring public management performance is always a challenging issue. Although use of Big Data opens many opportunities and provides insight recently unknown in public governance, it also carries a lot of risks related to neglecting new forms of data and opinion expressing means that don't correspond to the form and type of data usually assessed and analysed. For instance, performance indicators have diverse functions for different stakeholders over the life-cycle of a public policy, and the search for better indicators is an ongoing effort. However, activities could be challenged by the different interpretation of results of current performance measurement systems in open societies and competitive democracies (Johnsen 2005). Furthermore, Big Data analysis and the evidence it provides could be ignored in the development of effective and reliable policies in many policy areas affecting citizens. As a result, relevant social actors won't be engaged in the process of open policy-making. The ambition of this project was to advise on development of new types of evidence-informed and targeted policy design as well as to reflect on more elaborated enforcement and monitoring tools informed by data from various sources. Another goal was to facilitate interpretation of Big Data for more fluent public communication with clear understanding of legal, sociological,

cultural, political, economic and behavioural aspects. Ultimate goal beyond results of the project was for public administrations to have open and collaborative vision based on citizen's support and participation. It also had to be considered that dealing with Big Data in social sciences mean also dealing, at a great extent, with people's perceptions, life histories etc. Thus, for the years to come the combination of social sciences analysis and techniques and computational sciences to further develop and promote data content for policy driven implications must have in view inclusive and participatory societies (Mergel 2016).

WP2 aimed to identify the relationship between Big Data practices and policy development with the aim of enhancing the policy development process through Big Data analytics. The proposed approach was to identify the aspects of policy development that are affected by Big Data practices and analyse the associated benefits and risks. Further, to enhance the policy development process by incorporating the use of Big Data through the development of scalable and transferable methods.

The objectives of WPs were:

1. To identify the policy development aspects that can be influenced by Big Data practices;
2. To analyse potential benefits and risks of Big Data-driven policy development;
3. To define relationship between evidence-based policy development and citizens' participation;
4. To elaborate methodology for data-driven policy development with analysis of participatory elements included;
5. Define (scalable-transferable) methodology for policy development, which includes the following subtasks:

- policy making
- iterative policy modelling, testing and implementation;
- policy enforcement and compliance monitoring.

SMARTPolicy project proposal wasn't funded in the end due to high number of quality proposals and huge competition in that call.

6.2.4 SECURE DATA SHARING PRACTICES

Within recent years, researchers from various disciplines, such as economy, biology and political science, have increasingly seeked for access to confidential OS. Safe and secure data has actually become a strategic issue for archives in more than one way. According to the Administrative Data Research Network, 4 types of safety are most commonly discussed[42]:

- safe projects
- safe people
- safe, de-identified data
- secure environments

These have more recently also been supplemented with a fifth, "Safe analytic outcomes" The model of the "five safes" have gained increased actuality recently because of the GDPR idea of "built-in data protection". Such a situation was favored by the conjuncture of converging

[42] https://adrn.ac.uk/protecting-privacy/

elements: emergence of techniques to process huge amount amounts of data (e.g. longitudinal datasets issued by administrations); the creation of new and highly advanced statistical modeling tools; the opportunity to enhance datasets by matching one to another.

Yet anonymisation of very detailed data, whether medico-administrative data or tax data dealing with economic issues for instance, raises important dilemmas in terms of quality for statistical analysis perspectives. In turn, confidentially difficulties must be addressed. Even more so that the general public just as well as policy makers have repeatedly expressed concerns in respect with the ethical, legal and social issues involving the use of personal data. Related privacy issues have led to more specific EU regulations with the approval by European Parliament of the General Data Protection Regulation (GDPR), replacing the existing Data Protection Directive. This latest regulation is meant to update, harmonise and strengthen data protection law across Europe and beyond while allowing for exemptions of strict privacy regulation for the legitimate purpose of academic and applied research[43].

Questions regarding secure sharing data practices are tangled up with other contemporary obstacles amongst which we can cite, if only to pull up a few challenges: accessibility of data whether, when from a distance or when in situations of mobility; the use of datasets from different countries, leading new possibilities of longitudinal datasets, etc.

6.3 RESEARCHERS' NEEDS

In a matter of a few decades, data has become a major player in the scientific, institutional and corporate realms[44]. Official statistics has a very central role. This trend is related to strong political and citizen-based demand of expertise, dependent on the gathering of objective scientific data usually apprehended as evidence-based, empirically grounded knowledge (Struijs, Braaksma & Daas, 2014). This ongoing demand of expertise from data producers draws upon societal evolutions, characteristic of a risk society (Beck 1992). The decline of strong social institutions, often understood as traditions tied to family values, religious practices, gender roles or social classes paved the way to the massive production of scientific data as early as the 1950's; this process was accelerated in the late 1980's with the downfall of some of the main ideologies upholding societies in the Western world. Mainly, corporations and governments, seeking for new narratives and evidence to support decisions and sustain control, turned to social science and humanities for answers. But while this trend favoured the production of huge amounts of data, citizens and organisations are now more and more discontent with the current state of Official Statistical affairs. OS produced by NSIs is under criticism of neoliberal capitalist perspectives based on the general idea that statistics that are helping build tomorrow's economy should not be the by-product of governments bureaucracy[45], nor rely on "government's tired Economic Models" and related data[46]. This "statistical disillusion" must not be too quickly dismissed because it echoes with some of the

[43] http://www.3quarksdaily.com/3quarksdaily/2016/05/personal-data-for-the-publicgood.html#sthash.Lxla4cXQ.dpuf.
[44] Deliverable D8.4 (Final report proposing portal resource discovery functionality for a search/ browse portal interface)" http://www.dwbproject.org/about/deliverables.html Mike Priddy & Marion Wittenberg (DANS), "*What Researchers Want…From a Resource Discovery Service for OS Microdata*", 2nd European Data Access Forum, March 2015
[45] *Ibid.*
[46] *Ibid.*

researchers' complaints, preventing data sharing practices even further. Because, while civil society expresses dismay, researchers make similar claims, worsening a situation that still needs many encouragements.

It is well-known that social sciences have been suffering from a poor culture of data provision in comparison to other scientific fields, and of polysemy as to the boundaries on what constitutes social sciences research. The first is due to fragmentation of research and inadequate infrastructure; the second is due to the interdisciplinary nature of social sciences [47]. And while a large part of academics does not actually share their data and for those who do, their needs are changing. It is thus essential to better understand what this statistical disillusion feeds into and what elements keep researchers from sharing the data they produce. Likewise, for those who use confidential microdata, how have their needs changed and can DASs better serve their expectations. Studies conducted within DwB highlight a pattern of emerging needs and manners to improve data access and sharing practices.

6.3.1 Pilot study assessing researchers' needs in Slovenia

In order to explore the use of official statistics microdata and to recognise the practices and needs of researchers, four interviews with researchers that are using official statistics microdata at their work were conducted by ADP. We chose researchers that came from different research fields and are accessing and using data in slightly different way, which is reflected in their answers. The interviewees included an associate professor at the Centre of International Relations and an assistant professor at the centre for organisational and Human Resources (Both in the Faculty of Social Sciences, University of Ljubljana), a researcher at the Institute for Economic Research and a researcher at the National Examination Centre.

Interviews took place at the institutions where participants were employed and lasted for approximately half an hour. The interview protocol was an adapted version of the protocol that was used for other interviews in this work package (with researchers that use historical data). The main goal of the interviews was to identify possible issues regarding access, use, dissemination and quality of official statistics microdata as perceived and experienced by researchers.

Research areas in official statistics

Researchers that use official statistics microdata at their work named various research fields where they apply official statistics data. Among the fields mentioned were international trade and business, evaluation of economic and social programmes, labour market, employment and social policies, pre-schooling, education and vocational training, national examination and socioeconomic status, living conditions and welfare. Similarly, the data they use is as versatile as the research fields. Microdata they use are usually individual level data, although the researchers in the fields of economic use macro data (reports from official statistics or other administrative sources, SI-STAT data portal etc.), as well.

[47] Report: "Data collection strategies: CESSDA organisations and their relation to data collections outside CESSDA (D10.5a)", p.58. http://ppp.cessda.net/doc/D10.5a_Audit_collection_strategies.pdf

Three of the researchers actively use official statistics microdata, although only two of them on a regular basis. The National statistical office was named as the main source of microdata by all the participants, in particular Slovenian Census, European Union Statistics on Income and Living Conditions, Labour Force Survey, Slovenian Business Register, Community Innovation Survey and The survey on usage of information-communication technologies. Some of the researchers use other national or international official statistics/administrative data sources as well (e.g. Eurostat, OECD, IMF, ILO, Financial Administration of the Republic of Slovenia, Bank of Slovenia, Institute of Macroeconomic Analysis and Development, Agency of the Republic of Slovenia for Public Legal Records and Related Services). The microdata they use are both population (census) and survey based. Three of the researchers we interviewed collect their own microdata as well.

Accessing microdata

Generally, researchers reported no major issues regarding finding and accessing macro data as they are rather easily accessible and well presented. However, some remarked that the Statistical office of RS and Eurostat quite often change their website structure and thus access to data can be cumbersome until users adapt themselves to the new website and its content. Moreover, they also noticed that the national statistical provider does not always publish all the data on the aggregate level, but they are able to find more data (in terms of greater availability of variables) in Eurostat database.

On the other hand, accessing microdata poses some challenges. Two of the researchers accessed microdata through the secure room, located at the Statistical Office of RS, while others received microdata on portable devices, e.g. CD. None of the researchers reported having experience with remote access to official statistics data, mainly for the reason of statistical software – the remote access is supported only by STATA programme and our interviewees are either not familiar with this statistical package or they experienced some troubles with adjusting the right version of the programme. They also adverted that constant updates on both sides delay the work process.

The process of acquiring the access to microdata is a rather troublesome and the bureaucratic procedure can take a long time for the Data Protection Committee to make a decision.[48] But once the decision is made and access is granted, it seems things go smoother and researchers claim that statistical office is very helpful and flexible, they are responsive and willing to explain and support the researchers. However, the length of the process was named as a drawback by some of the researchers, one of the explicitly claiming that he does not use the microdata anymore only for the reason of bureaucracy.

The researchers that use SURS's secure room reported positive experiences although they pointed out that there are some limitations (e.g. export of the data, opening times of the secure room), but they consider them as a trade-off to access microdata and therefore they try to adapt to the situation.

[48] SURS - ACCESS TO MICRODATA FOR RESEARCHERS HTTP://WWW.STAT.SI/STATWEBARHIV/EN/MAINNAVIGATION/DATA/FOR-RESEARCHERS/GENERAL

Quality of official statistics microdata

All the researchers we interviewed evaluated the quality of official statistics data as good quality, although they sometimes came across some ambiguities. In response, the statistical office reacted quickly to correct them. However, two of the researchers remarked that they would like more detailed documentation on methodology and metadata (e.g. how is the data collected, procedures on missing data), while the researcher working mainly with business data believed that the documentation is adequate and well explained.

Dissemination of official statistics data

All of the interviewees reported that they attend scientific conferences where they present their research results. Statistical office also requests that researchers report the outcomes of their research work on official statistics data, e.g. list the papers they publish. The list of published papers is also distributed through the Faculty's communication channels. Two of the researchers are also teaching at the Faculty of Social Sciences and one of them includes official statistic aggregate data into his lectures, while the other one expressed some reluctance towards that as he believed that the data he uses in his research might be too difficult for students. Researchers often liaise among themselves and share their experiences about data research, although these collaborations are mainly established in small groups within the department or the institution.

When asked about the role of the archive in the dissemination of data, participants gave rather opposite opinions. While one of them claimed that the data that the archive offers has rather historical than current research value and that as such is not useful for her research as it is not updated. The other three interviewees expressed more favourable thoughts, although it was also pointed out that the datasets that the archive offers might be more useful for students than for researchers, due to the anonymisation procedures. One of the researchers noted that longitudinal analysis is playing a greater role in research nowadays and in this regard, the archive's role is crucial to facilitate the process of accessing relevant datasets. The other researcher pointed out that we still need a stronger presence of archives on national level.

Future

When talking about research interests in the near future, researchers share the opinion that it is always better to have more data available. They believe that more microdata will be required, especially as the techniques for analysing the data develop and we are able to gain deeper insights. They notice that there is more data collected than actually available and they believe it will be beneficial to allow access to more data collections. Especially in Slovenia, due to the small size of the country, it is easier to collect more data. However, as one of the researchers pointed out, there is also the issue of personal data protection to be addressed in this regard. Researcher who works with business data, also remarked that the hopes that the availability of data will not reduce in the future, although he acknowledged the issue of burden for businesses to report the data.

6.3.2 TRANS-NATIONAL COMPARABILITY

"Comparability is important. For cross-national research, you have to know exactly how a concept in measured in the different countries."

Multiple languages, poor knowledge of national contexts and heterogeneous metadata information throughout Europe is one of the first obstacles mentioned by academics in another study about researchers' needs. These issues can quickly become barriers to exploit and interpret results, or simply compare similar datasets, and hence conduct research[49]. Above and beyond these bottlenecks, transnational collaborative work is still difficult. As it is brought forward by the OECD Expert Group for International collaboration on microdata access[50], most countries allow transnational access to their confidential microdata, though in different ways and at varying degrees. Researchers may have to travel to the secured data centre onsite rather than access remotely but it is still possible. Simply cross-national collaborative research is seriously compromised as soon as datasets are held in several centres in different countries, attempts to combine datasets and analysis is lead by comparing outputs. In general, working within European academic European networks handling OS implies dealing with a series of setbacks: obtaining clear and comparable data documentation, gaining accreditation to access data through various procedures, managing different modes of access and output checking, technical settings, etc.[51].

Recommendations:

- The use of a multilingual thesaurus would overcome the language problem.
- One needs to have a basic kind of knowledge about the national context of a country to be able to interpret the results of the analysis.
- Reports with key indicators are essential in this respect.[52]
- The development of e-infrastructure, developed as a type of an independent secure access system, could allow researchers located in different places/countries to work together across borders with confidential microdata held in different institutions/countries, as each.

DwB D8.4 suggests that to gain in comparability possibilities, researchers could benefit from contextual information, identifying certain variables. Associated literature, citations and other documentation related to a OS dataset of interest could be made visible to the research when linked to a project-specific Virtual Research Environment (VRE). In a similar manner linking and annotating resources (both metadata and imported documents) should be made possible.

[49] Deliverable D8.4 (Final report proposing portal resource discovery functionality for a search/ browse portal interface)" http://www.dwbproject.org/about/deliverables.html Mike Priddy & Marion Wittenberg (DANS), "*What Researchers Want…From a Resource Discovery Service for OS Microdata*", 2nd European Data Access Forum, March 2015.

[50] https://www.oecd.org/std/microdata-access-executive-summary-OECD-2014.pdf

[51] Grenet J., European Data Access Forum, March 2012, Luxembourg.

[52] http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d8_4_portal-resource-discovery-functionality_final-report.pdf

6.3.3 ACCESSIBILITY AND REMOTE EXECUTION

Accessibility is another important dilemma that needs to be alleviated to help academics exploit data centres and/or databases efficiently. This is even more true in the present context where the number of available datasets had greatly increased. In parallel, there is an even greater variety of data holders and providers. The landscape has become increasingly fragmented for discovery of data, with insufficient and non-harmonised metadata, most often not translated in English[53].

- **Fix data centres: a concept for the past?**

Up until the past ten years, data access was only given to a limited number of scientific candidates who accepted to go to onsite secures centres on site, usually in National Statistical Institutes or some statistical departments in the ministries, handling data sometimes under the supervision of the data producers. Accessibility issues reported by researchers. This previous situation is quickly evolving. Remote data centres have been blooming, points of access have been multiplying, all the while the quantity of data has been growing.

- **Promises and challenges raised by the open data movement**

The "open data" movement also strongly affects the situation as there is growing demand for public use files, which can potentially be disseminated widely through the web. Hypothetically, "*public use files enhance democracy (making information available to a large public) and strengthen statistical literacy (allowing students and beginners to perform analyses and improve their skills)*"[54]. But this latter development is not without its' own challenges, because to achieve open data, "*countries must strike a balance between the opposing needs of, on the one hand, keeping as much information as possible in the file to make it useful, and on the other hand, preventing any form of re-identification of statistical units. This is one reason why few countries produce public use files at all, and production is limited even in the countries that do*[55]".

- **Accreditation: a lengthy process**

Another accessibility difficulty is the duration of some accreditation processes. As revealed in DwB:

- Too long of a wait for accreditation
- Several steps between accreditation and real access
- Outputs checking are too long for 1/3 of the researchers

[53] Deliverable D8.4 (Final report proposing portal resource discovery functionality for a search/ browse portal interface)" http://www.dwbproject.org/about/deliverables.html Mike Priddy & Marion Wittenberg (DANS), "*What Researchers Want…From a Resource Discovery Service for OS Microdata*", 2nd European Data Access Forum, March 2015

[54] http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d3-1_researchers-accreditation_report_final.pdf p. 24

[55] *Ibid.*

- **Eliminate any distance to access point: remote access?**

In case of short term mobility, a researcher is unable to resume data processing instantly in order to modify or test new models or refine the analysis; In case of long-term mobility, researchers have to ask for a new accreditation (Cornuau & Silberman, 2015).

- Time consuming
- Financial cost
- organisational problem
- No Interaction or discussion between researchers
- No usual documentation at hand
- Discriminating

Recommendations:

- Develop remote access systems and remote execution
- Reduce waiting time throughout the accreditation process[56]
- Prevent obstacles to accessing research, whether restrictions are due to personal computer constraints (Cornuau & Silberman, 2015) or situations of mobility
- Information about availability, access conditions and procedures should be made available at the appropriate places on a national level (fact sheets from DwB exist but are not distributed or exploited on a national level by CESSDA's SPs or NSIs
- Should be possible to find historic datasets & studies.
- Must be automatically identified, updated and maintained
- If an issue is spotted, it should be possible to report it to the data provider/ producer.

6.3.4 User-generated and social networking oriented content

"It would be great if there were a possibility to share the work on harmonisation. …user groups around specific data or specific topics. These user groups share expert knowledge, papers, sometimes even syntax of the analyses. It would like to have the possibility to add comments. This can also help other researchers."

One of the findings brought forward by DwB (D8_4) relates to researchers will to share information in small groups of specialists (e.g. users sharing the same datasets) and build communities of knowledge and expertise. Such gatherings could be the opportunity to create user generated content: participants could annotate their search results and share these with their colleagues.

Recommendation: **sharing, collaborating and building communities via a VRE (virtual research environment)**[57]

[56] Several researchers stated the fact they couldn't access their own research by the means of their personal computer because of the following obstacles: need for specific software not provided by the secure centre; transfer of programs can take time; communication with other colleagues is difficult; no web access (Frédérique Cornuau, Roxane Silberman, Researchers' needs : understand how they work to implement a EU-RAN, EDAF, Luxembourg, 24th – 25th of March 2015). See also http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d8_4_portal-resource-discovery-functionality_final-report.pdf

[57] DwB, D4_8 http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d8_4_portal-resource-discovery-functionality_final-report.pdf p.42

- Communities should be able to build their own specialised portal
- Communication & social media tools could help build knowledge bases & communities
- Metadata enrichment by the researchers could take place?
- Aid to set up user groups, Wikipedia on OS data and crowd sourcing of expertise.[58]
- More elaborate user-generated and social networking oriented content could include sharing documentation, work on harmonisation, expert knowledge, literature, papers, and sometimes even syntax of the analyses.

6.3.5 DATA DOCUMENTATION AND ADVANCED RESOURCES DISCOVERY

Complex and comparison searches have been pointed out by scientists as important challenges ahead (Priddy & Wittenberg, 2015[59]), especially since incomplete sets of metadata can obstruct meaningful comparisons. Documentation or metadata about data allow resource discovery and understanding of the data. Yet, at the moment, aspects of the data collection (sampling procedure, mode of collection, trend breaks, changes over time in methodology, data collection mode, question wordings etc.), weighting of the data and other methodology characteristics are often lacking; NSI's are only publishing the documentation of recent surveys. More extensive documentation reports are needed.

- "Ideal documentation" depends on particular research interests, even if what is today ideal might be very limited in the future, or too much detailed for a very basic type of research.
- Especially important the ability to search for a wide spread of geographical and temporal variables.
- about the field workers (sex, age, previous experience as a field worker), in order to allow a detailed investigation of the data quality and possibly of some systematic errors[60]

Recommendation: Need for high quality, consistent & citable metadata

High-quality machine-readable metadata could enhance the findability of OS datasets. DwB D8.4 recommends the possibility for researchers to select information provided by a search query and/or browse this information, in such a manner that the researcher can store the query itself and its results for future use or sharing[61]. It should additionally be possible to cite (via a persistent identifier (PID) a metadata record (or set of records), as citation of metadata record would be crucial for scholarly publication. Other recommendations brought were:

- Need for high-quality & extensive metadata
- Consistency in metadata available & changes clearly identified.
- Questionnaires, codebooks, complete methodological documentation and more…

[58] http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d8_4_portal-resource-discovery-functionality_final-report.pdf
[59] Deliverable D8.4 (Final report proposing portal resource discovery functionality for a search/ browse portal interface)" http://www.dwbproject.org/about/deliverables.html Mike Priddy & Marion Wittenberg (DANS), "*What Researchers Want…From a Resource Discovery Service for OS Microdata*", 2nd European Data Access Forum, March 2015
[60] http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d8_4_portal-resource-discovery-functionality_final-report.pdf
[61] DwB,D4_8 http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d8_4_portal-resource-discovery-functionality_final-report.pdf p.42

- Improvement of the quality of data documentation and support (Cornuau & Silberman 2015).
    - Poor quality of documentation is an issue ("clear and up-to-date documentation, standard format for standard variables (e.g. time variables, geographical units id, etc.)
    - Need for a reactive human support (e.g. hotlines)
- Language issue is particularly important too: which language(s) for documentation?
- Create the possibility to search or browse on 3 dimensions: topics, time, and location.
- Possibility to search for variables, which is not possible at the moment with many NSIs.
- Have access to a quick overview of all resources by country, year, topic, accessibility and type of data[62]

To sum up, we see that researchers want to work faster, anywhere, with their own personal computer in their workspace, in a collaborative way.

6.4 LANDSCAPE OF OS: IDENTIFYING DATA ACCESS, AGREEMENTS AND RELATED STRATEGIES

As it was established during the PPP project, it is important for CESSDA to have a clear view of the landscape via a series of criteria, in particular:

- identifying data that is not currently accessible;
- listing strategies for acquisition policies;
- harmonising dissemination policies;
- realistic planning for networking with data agents at various phases in the life-cycle of data;
- strengthening the bonds among research actors (i.e. academic-administrative-business actors);
- internationalizing research in terms of provision, while focusing at national/cultural boundaries in terms of production[63].

Similarly, as for as other types of data the OS data services challenges can be summarised under three broad questions:
1. How far is social sciences research production covered by data archives the Country members of CESSDA?
2. What is produced, how is it provided and where – if not through CESSDA?
3. What can CESSDA do to enhance acquisition and dissemination of the collections existing in the European research area?[64]

In the case of OS, much information was collected within the scope of CESSDA research efforts. This is especially true when it comes down to establishing what type of data is provided in NSIs. Even if past efforts have made great progress, major blind spots remain to determine what

[62] http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d8_4_portal-resource-discovery-functionality_final-report.pdf
[63] Report: "Data collection strategies: CESSDA organisations and their relation to data collections outside CESSDA (D10.5a)", p.59
[64] *Ibid.* p. 59-60

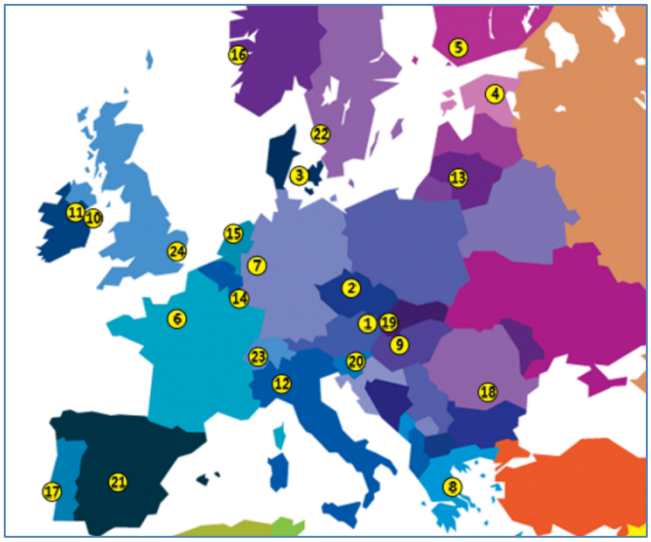
datasets are outside of CESSDA's holding even though several initiatives have tried to map the situation.

6.4.1 TRANSNATIONAL PATHWAYS TO ACCESS OS VERSUS NATIONAL LEVEL BOTTLENECKS

- **Previous efforts to identify datasets throughout Europe**

Much effort has been deployed to discern data providers, data producers and/or data archives throughout Europe, may it be initiatives from the Official Statistics office of the UN, major archives like the UKDA[65], or in the frame of research conducted within EC projects. Even more in the case of OS, where effort has been put into identifying data holdings; initiatives have been carried to create proper lists of national representatives, either NSIs or statistical departments of ministries. First results showcase the following map of data centres working on OS:

Picture 1: Map of data centres providing an access to OS[66]



More importantly, DwB provided fact sheets about how to access some of the main archives holding statistical data here:

http://www.dwbproject.org/access/accreditation_db.html

However, the growing number of data producers, the always vaster points of access for the same set of data, the support of new technologies in generating and holding data, etc. leads to the dispersion of data centres. This moving landscape complicates a situation filled with obstacles already present on a national level.

[65] https://www.ukdataservice.ac.uk/get-data/other-providers/data-archives/europe
[66] https://www.ukdataservice.ac.uk/get-data/other-providers/data-archives/europe

- **Uneven playing field of official statistics at a national level leads previous studies to focusing on NSIs for comparability purposes**

Access to European datasets is rather well organised but there is a lack of homogeneity at a national level. The gap between the European landscape and national settings is growing as recent improvement in accessing official data throughout Europe has been concluding. **Highly detailed data is yet very difficult to access[67], given** legal limitations for data to cross borders, the existence of multiple accreditation processes, the diversity in modes of access, infrastructures and information. Progress is nonetheless widespread. Trans-national access is becoming a reality, with the help of European (Eurostat's) data.

- **Scattered sources of data on a national level raises data access and visibility issues**

Yet many projects need national data. But the problem on the national scale is the scattered nature of datasets. In addition, multiplication of points of access as the French case emphasises.

French case: as the web investigation highlights, France hosts dozens of data centres with no centralised platform or infrastructure bringing together all the different data available in the field of official statistics in such a manner that researchers are unaware of the resources at their disposal. This situation is also directly related to the evolution of the IT landscape: the proliferation of different access points for the same datasets[68] lead to overall confusion regarding where the same data is deposited and who holds what types of datasets. The lack of a centralised infrastructure or a common platform weakens the re-use of data.

Recommendations

To overcome this obstacle, the OECD Global Science Forum report on data and research infrastructure for the social sciences recommends that Official statistics and research users: Mechanisms should be found to bridge across the communities of official statisticians and social scientific researchers[69]. This global solution must compose with the harsh reality on local legislation and difficulties in implementing national changes in regards with the complex of laws, backlash of general public and watchdogs and burdensome of legal procedures and such changes present financial and staff restrictions put at risk even existing outputs on NSI[70].

6.4.2 FROM DATA ARCHIVED BY NATIONAL ORGANISMS TO THE PORTION OF DATA AVAILABLE

To have a better understanding of potential data accessed through data, it is also necessary to know what data centres have in their holding. Hereafter, we present finding from DwB project about types of data files in NSIs across Europe and highlight where these data files come from.

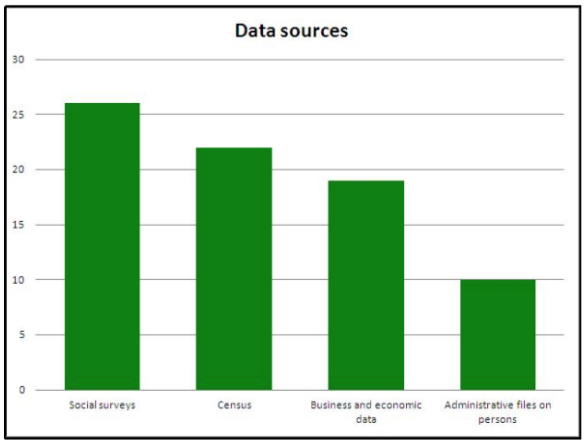
[67] http://www.dwbproject.org/export/sites/default/edaf0/dwb_edaf_session-b_tna_accreditation.pdf
[68] http://renatis.cnrs.fr/IMG/pdf/SILBERMAN_08102013.pdf
[69] http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf
[70] http://www.dwbproject.org/export/sites/default/edaf0/dwb_edaf_session-h_access_across_borders.pdf

- **Which data is released to researchers?**

What are the sources of data released to researchers? Social surveys? Censuses? Business or economic data? Or is administrative data?

If we have a closer look at the figure hereafter on the data sources stored and released to researchers, we can note that social surveys are most commonly used; business and economic data follows. A higher disclosure risk, in smaller countries in particular, could however limit access to the latter data. Paola Tubaro underlines the fact that census data, while still disclosed and widely used, knows greater cross-country variation in comparison with the other data sources. Finally, administrative data most commonly relies on in Nordic countries, even though tendency of storage and use seems to be developing in other parts of Europe[71].

Picture 2: Frequencies of Data Sources



- **How do researchers access these files?**

Modes of access:

- Transmission to researcher (SUF and equivalent) is the most common mode;
- Secure modes of access, both onsite and remote, are gaining ground (remote access more than remote execution)
- Availability of PUFs is also growing, though at a slower pace;

In some countries, data archives disseminate data on behalf of NSIs —including sometimes secure systems (France, UK).

[71] DwB D3.3 Researcher accreditation - current practice, essential features, and a future standard p. 4. http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d3-1_researchers-accreditation_report_final.pdf

Picture 3: Frequencies of "Modes of access"

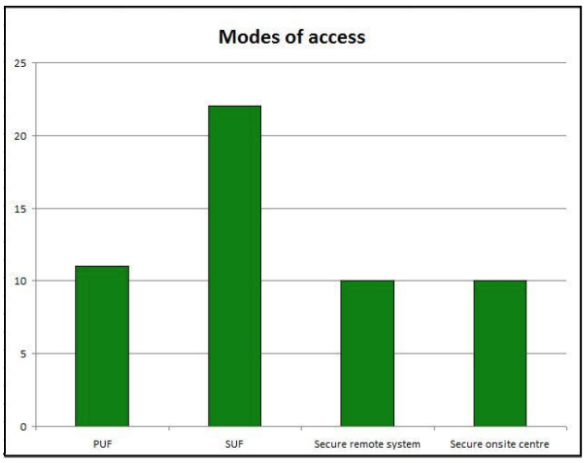
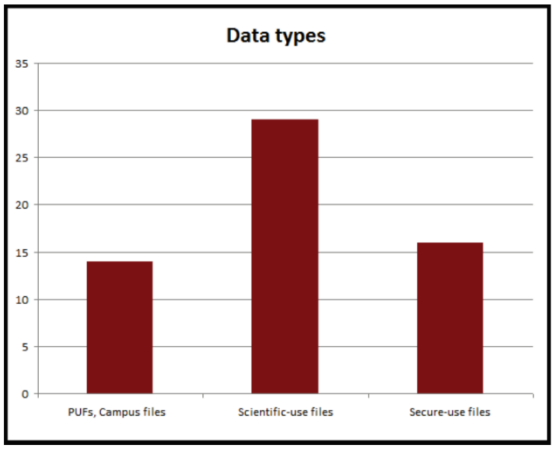


Figure: Modes of access, absolute frequency. PUFs, Public Use Files: widely available files, possibly freely downloadable from the Web. SUFs, Scientific Use Files: files intended for research use and transmitted to the researcher. Secure remote systems: include both remote execution and remote access; the researcher connects remotely to a secure environment under NSI's control. Secure onsite centre: Data laboratory, often on NSI premises, where access is controlled.

- **What types of data are currently released to researchers?**

In this figure, we can examine the number of NSIs that provide secure use files[72] (SUFs). Secure use files: Scientific use files[73] (SUFs); Public use files (PUFs), Campus files[74].

Picture 4: Frequencies of "Data types"



If scientific use appears to still have the upper hand remaining the primary mode of dissemination of microdata for research purposes, open data is shuffling the deck. The growing demand for public use files, potentially available to the general public on the Internet, are still

[72] Data from which direct identifiers have been removed, but to which no further methods of statistical disclosure control have been applied.
[73] Data without direct identifiers to which methods of statistical disclosure control have been applied to reduce disclosure risk.
[74] Data that have been subject to heavy statistical disclosure control methods that have eliminated almost all disclosure risk. Campus files are not reliable for substantive analyses and can only be used for teaching.

however of poor quality for the moment; too heavily anonymised, they prevent fine scientific analysis.

6.4.3 TYPE OF AGREEMENTS WITH PRODUCERS IMPACTING COVERAGE

Table 13: Data access conditions for NSIs

Data access policy	Archive/ Institution/ Project	Country
General Conditions Access to anonymised non-personal data is granted for scientific research purposes. Data is non-personal if the data subject cannot be identified through reasonable use of means. Users must be affiliated to recognised higher education or research institutions. **Conditions for Students** Access is granted to anonymised non-personal data. Legal Framework Bundesgesetz über die Bundesstatistik (Bundesstatistikgesetz 2000)	Statistics Austria	**Austria**
General Conditions Access to anonymised data can be granted for scientific research. Beneficiaries include, but are not limited to, researchers affiliated to higher education or research institutions. **Conditions for Non-Resident Researchers** Same conditions as national researchers. Transmission of data is possible only to "safe countries" that comply with European personal data protection legislation, or offer equivalent protection. **Conditions for Students** Data are usually not provided to students Legal Framework Statistical law of 4 July 1962, modified 22 March 2006.	Statistics Belgium	**Belgium**
General Conditions Individual anonymous data can be provided for the purposes of scientific work to higher education or scientific research institutions, with the permission of the President of the National Statistical Institute. Legal Framework Law on Statistics (Promulgated SG 57/25.06.1999, amended 2010)	National Statistical Institute	**Bulgaria**
General Conditions Producers of official statistics may, on the basis of a written request, provide individual statistical data without identifier for the purpose of performing the activities of scientific research. Applications for research access to data shall state the purpose of the use of the statistical data. A contract shall be signed, according to which the user shall be held financially and criminally responsible to use statistical data only for the purpose stated in the request, and shall not provide these data for inspection or use to unauthorised persons, and shall destroy such data after use. Producers of official statistics shall keep records of research usage of the data. Contract violations are punished with a fine. Legal Framework, The Official Statistics Act (the Official Gazette Nos. 103/03, 75/09 and 59/12)	Croatian Bureau of Statistics	**Croatia**
General Conditions CYSTAT may release microdata for the sole use of scientific research, provided researchers' applications are approved by its Confidentiality Committee. When applications are approved, microdata may be released after an anonymisation process which ensures no direct identification of the statistical units, while still maintaining usability of the data. **Conditions for Non-Resident Researcher**s There are no special conditions in place for non-resident researchers. All applicants (resident or non-resident) are treated the same way. **Conditions for Students** Postgraduate students may apply via their supervising professor or the head of their department. A permanent member of the academic staff in an identifiable entity of a higher education institution (such as faculty, school, department, research institute, etc) needs to be the applicant. Legal Framework Statistics Law No. 15 (I) of 2000	CYSTAT	**Cyprus**

General Conditions Data without direct identifiers can be provided by CZO for scientific research purposes. **Conditions for Non-Resident Researchers** Same conditions as national researchers **Conditions for Students** Because data can only be provided for scientific research purposes, and final theses are not considered as research, micro data cannot normally be provided to students. Exceptions are possible when the student participates in a research project. Legal Framework Act No. 89/1995 Sb, Act No. 101/2000 Sb	CZSO	**Czech Republic**
General Conditions Researchers affiliated with pre-approved Danish research institutions can be granted authorisation to access register data, collected from the 1970s to the present. Access is given to data at personal level (individual persons or firms) for several years, albeit stripped of direct identifiers. Access is given on the basis of a need-to-know principle. Users can be authorised to link register data with data from other sources, such as surveys. **Conditions for Non-Resident Researchers** Non-resident researchers can get access to microdata through an affiliation to a Danish pre-approved research institution, which takes responsibility for use. What matters is affiliation (that is, the institution that takes responsibility and can be sued in case of breach) rather than physical location of the user (indeed authorised Danish researchers can connect to the system from abroad). Legal Framework Statistical code, 2006. For further details click [Here] (in Danish)	Statistics Denmark	**Denmark**
General Conditions Statistics Estonia can disseminate microdata (including confidential microdata) for scientific research purposes. Microdata may be accessed by legal persons or agencies, not freelance natural persons, after approval of a written application by the Confidentiality Committee (an internal body of Statistics Estonia), signature of an agreement with the legal person having submitted the application, and signature of a confidentiality pledge by all researchers involved in the project. **Conditions for Non-Resident Researchers** Non-resident researchers may have access under the same conditions as national researchers. **Conditions for Students Master's** and Doctoral students may have access under the same conditions as confirmed researchers. Legal Framework Official Statistics Act of 2010	Statistics Estonia	**Estonia**
General Conditions Statistics Finland provides microdata for research use. Data with possibility for indirect identification are available only for secure remote access or use at the onsite safe centre (Research Laboratory). Release of data outside Statistics Finland (often, tailor-made extractions of samples from existing data) is possible only for data protected against direct and indirect identification. These data are supplied to researchers on CD or DVD or on a memory stick. **Conditions for Non-Resident Researchers** Anonymised data, without possibility for direct or indirect identification, can be sent abroad. Researchers within EU have the possibility to gain access to data from which only direct identifiers have been omitted, through remote access or by visiting the onsite Research Laboratory. The researcher should have contacts with a Finnish research organisation. **Conditions for Students** Students can obtain microdata for statistical analyses. Legal Framework Statistics Act 280/2004; Personal Data Act 523/1999; Amendment of the Personal Data Act 986/2000.	Statistics Finland	**Finland**
General Conditions Fully anonymised microdata (public use files) can be accessed freely by any user. More detailed versions of data can be accessed for research purposes, a condition that is interpreted as involving the production of new knowledge of general significance whose results are to be published. Use for commercial purposes is proscribed. **Conditions for Non-Resident Researchers** Non-resident researchers may have access under the same conditions as national researchers. However, for confidential data distributed under secure access, application requires additional supporting documentation. **Conditions for Students** Master's students, PhD students, and post-doctoral post holders, under supervision of a tutor, can access data under the same conditions as	INSEE	**France**

confirmed researchers. Legal Framework Law on Statistics, 1951; Data Protection Act of 1978; Archives Act		
General Conditions Data can be accessed by researchers from recognised research and higher education institutions based in Germany. **Conditions for Non-Resident Researchers** Non-resident researchers may access public use files, and may use scientific and secure use files at the safe centres of the statistical office (Forschungsdatenzentrum, FDZ). They cannot receive delivery of scientific use files outside Germany. **Conditions for Students** PhD students based in German Universities are eligible at the same conditions as confirmed researchers. All other students are eligible at more restrictive conditions (access to a maximum of five data sets; no bespoke data preparation). Legal Framework Statistical law, BStatG 1987	DESTATIS	**Germany**
General Conditions Access to data can be granted for non-commercial purposes only. Four modes of access are available, distinguished by degree of anonymity of the data and terms of data use. (1) Campus files are heavily anonymised data for use in teaching, but not suitable for research. For substantive research, (2) scientific use files (de facto anonymised datasets) are available via secure download. Secure IT solutions enable the use of more detailed, weakly anonymised data sets via (3) onsite use or (4) remote execution. **Conditions for Non-Resident Researchers** Non-resident researchers can use the same access ways as domestic research. **Conditions for Students** Same conditions as national and international researchers (with support of Faculty member of their institution). Legal Framework German Social Code Book	IAB	**Germany**
General Conditions ELSTAT releases fully anonymised microdata from statistical surveys, not allowing direct or indirect identification of statistical units. It may also grant access for scientific purposes to data that enable the indirect identification of statistical units, provided: (a) an appropriate request together with a detailed research proposal are submitted; (b) the research proposal indicates in sufficient detail the set of data to be accessed, the methods of analysis, and the time needed for the research; (c) a contract specifying conditions for access, obligations of the researchers, measures for respecting the confidentiality of statistical data and sanctions in case of breach, is signed by the individual researchers, their institution (or the organisation commissioning the research), and ELSTAT. **Conditions for Non-Resident Researchers** Non-resident researchers may have access under the same conditions as national researchers. **Conditions for Students** There are no special conditions for students. Legal Framework Observance of statistical confidentiality by the Hellenic Statistical Authority (ELSTAT) is regulated by articles 7, 8 and 9 of the Statistical Law 3832/2010, by Articles 8, 10 and 11(2) of the Regulation on Statistical Obligations of the agencies of the Hellenic Statistical System (ELSS) and by Articles 10 and 15 of the Regulation on the Operation and Administration of ELSTAT.	ELSTAT	**Greece**
General Conditions In order to fulfil data requests for not public statistical information, the HCSO offers 6 data access channels for the users. 2 of these channels are available for all users without restrictions while 4 of them are available for researchers for scientific purposes only. For data access channels available for scientific purposes only, a researcher accreditation procedure is applicable and a contract and a confidentiality commitment must be signed. The following 2 data access channels are available for all users without restrictions: Tabular data (both predefined and customised) and highly anonymised Public Use Files are available to all users with the acceptance of Terms of Use. Requests for tabular data: Using this data access channel, access to tabular data can be requested. Following the positive evaluation of the data request from professional and data protection point of view, the requested tabular data will be sent to the person requesting the data using the preferred transmission mode (e-mail, post). Access to Public Use Files: Publicly available microdata sets with strong statistical	HCSO	**Hungary**

disclosure control can be accessed by using this data access channel for testing, teaching and research purposes. Data access channels available for scientific purposes only. The following 4 data access channels are available only for researchers for scientific purposes. The HCSO performs a researcher accreditation procedure for all data requests for these 4 data access channels. Release of anonymised microdata sets where direct identifiers are removed and further statistical disclosure control methods have been applied, as well as access to de-identified, but more detailed data in the Safe Centre or through remote access or remote execution, are reserved to researchers only. Safe Centre access: The HCSO offers access to de-identified microdata sets for scientific purposes in the safe environment of the Safe Centre operated by the HCSO in Budapest. Remote access: This data access channel offers access to de-identified microdata sets for scientific purposes in the safe environment of the remote access points operated by the HCSO under the same access conditions as the Safe Centre. Remote execution: For scientific purposes, the HCSO produces the requested research outputs inside its own safe environment based on the specifications/syntax files provided by the researcher. Release of anonymised microdata sets: By using this data access channel the HCSO provides anonymised microdata sets for the researchers for scientific purposes. Anonymised microdata sets are customised and only the variables requested are provided to the researcher. The secure access modes (Safe Centre access, remote access and remote execution) provide access to both ready-made and customised microdata sets. **Conditions for Non-Resident Researchers** Non-resident researchers (not limited to EU) may have access under the same conditions as national researchers. **Conditions for Students** Master's and PhD students may have access under the same conditions as confirmed researchers. Legal Framework Hungarian Act on Statistics: Act No. XLVI of 1993 / Right of Informational Self-Determination and on Freedom of Information ("Privacy Act"): Act No. CXII of 2011		
General Conditions Anonymised microdata files (AMF) are available for non-commercial research purposes. For scientific use files (called Research Data Files, RDF), researcher status must be proven. **Conditions for Non-Resident Researchers** Non-resident researchers can obtain AMFs. Access to RDFs is determined by the Statistics Act 1993. As this is national legislation, access can only be granted within the state boundaries. There are no restrictions on nationality provided the information is accessed within the state boundaries. **Conditions for Students** Post-graduate students are eligible (no undergraduates) Legal Framework Statistics Act, 1993. AMFs: Section 34 of the Act. RMFs: Sections 20(c), 32, 33, 38, 39, 42(1), 42(2) and 44.	CSO	**Ireland**
General Conditions Microdata files derived from Istat's surveys, are released free of charge and in compliance with the principles of statistical secrecy and protection of personal data. There are three types of highly anonymised data: two types of scientific use files (files for research purposes and standard files) and a public use file ("mIcro.STAT file"). Files for research purposes are developed in relation to statistical surveys regarding individuals and households as well as enterprises, and are created specifically for the purposes of scientific research. They are the most detailed files and they may be requested exclusively by: a) subjects belonging to Italian universities or research bodies and institutions to which the "Ethical code for the processing of statistical data outside Sistan" (a data processing regulation that concerns universities and other institutions outside the national statistical system) applies; b) other subjects that meet Eurostat requirements for the provision of microdata files in accordance with European legislation. Standard files from surveys on individuals and households may be accessed by a variety of users, but are restricted to study and research purposes. They are issued upon request with a valid reason for research or study purposes. Public use mIcro.STAT files are developed for some surveys starting from the relative file for research purposes, as a subsample. They contain a lower level of	ISTAT	**Italy**

detail in comparison with files for research purposes. **Researchers** (defined as those affiliated to, or fellows of, Italian universities or research bodies and institutions to which the above-mentioned "Ethical code" applies) can also access secure use files using Istat's onsite safe research data centre, ADELE (Analisi di Dati ELEmentari). **Conditions for Non-Resident Researchers** Non-resident researchers may have access under the same conditions as national researchers. **Conditions for Students There** are no constraints for public use files as they can be freely downloaded directly from Istat's website. To acquire such files, it is necessary to register at the area of the Istat website dedicated to them and to accept the terms of use. Access to other files is restricted, upon suitable application, to PhD students of institutions that are bound by the Ethical code for the processing of statistical data outside the National Statistical System (Sistan). Legal Framework Legislative Decree 322/1989 establishing the National Statistical System (Sistan); Ethical code for the processing of statistical data outside the National Statistical System (Sistan), 14/08/2004, n. 190		
General Conditions CSB provides two types of anonymised respondent-level data from social surveys (no business data). A small set of highly anonymised data from Labour Force Surveys is available online and can be used for teaching purposes. Other anonymised individual data are made available exclusively for scientific purposes, upon application and signature of a contract, and under the condition that results of research shall be published and have potential to benefit society as a whole. No other uses are authorised for these data. CSB also provides more detailed data for research, from which direct identifiers have been removed, through its secure remote access facility. **Conditions for Non-Resident Researchers** Non-resident researchers may have access under the same conditions as national researchers. **Conditions for Students** Post-graduate and Doctoral students writing a dissertation / thesis may have access on a need-to-know basis. Legal Framework Law of Statistics (1997, last amendments 2009) By-Law of the Central Statistical Bureau of Latvia, Cabinet Regulation 994, 2004	Central Statistical Bureau of Latvia	**Latvia**
General Conditions Respondent-level statistical data can be made available in two forms. Highly anonymised data from some social surveys, that allow neither direct nor indirect re- identification of statistical units (public use files), are available to all for use. Other microdata, with direct identifiers removed, are available exclusively for scientific purposes; only researchers with an employment contract with higher education or research institutions as identified in Lithuanian law, can apply. Statistics Lithuania currently carries out a project which will allow scientific institutions to work with microdata via secure Internet connection. After certain procedures (checking of the institution, signing of a contract, etc.), the microdata requested are prepared and uploaded to a special server. Users can receive only tabular results, previously inspected by the specialists of Statistics Lithuania. **Conditions for Non-Resident Researchers** Non-resident researchers, whose institutions qualify as Higher Education or Research Institutions, may have access under the same conditions as national researchers. **Conditions for Students** Doctoral students preparing their thesis may have access under the same conditions as confirmed researchers; undergraduate and postgraduate (master's) students are allowed to access only public use files and cannot apply for research-use data. Legal Framework Law on Statistics, 1993 – No I-270 (last amended 2010). Law on the right to obtain information from state and municipal institutions, 2000 - No 10-236 (last amended 2005) Lithuanian Science and Studies Act, 1991 - No. 7-191 (last amended 2002)	Statistics Lithuania	**Lithuania**
General Conditions For scientific purposes, STATEC may grant access to confidential data on its premises. The admissibility of the request for and the authorisation of access to the confidential data for scientific purposes are subject to the assessment of the merits and the scientific interest of the research projects for which the authorisation is requested, and also to the assessment of the scientific qualification of the applicant(s). The terms and conditions of access are determined	STATEC	**Luxembourg**

by STATEC. The studies and results of the research that are likely to be published or disseminated are checked by STATEC to avoid the disclosure of confidential data. Information that can lead to identification of a statistical unit cannot, under any circumstances, be disclosed. Legal Framework Law of 10 July 2011 on the organisation of the National Institute for Statistics and Economic Studies		
General Conditions Microdata access is only granted under strict conditions to a selected number of institutions or persons accredited as "research entities" or "researchers" for use in research projects. To be accredited, applicants have to demonstrate their knowledge and experience for handling potentially disclosive personal information. They also have to provide evidence that illustrates professionalism and technical competence. They have to demonstrate a commitment to protecting and maintaining the confidentiality of the data. **Conditions for Non-Resident Researchers** Conditions under which access to confidential data is granted is identical for Maltese and non-Maltese residents. **Conditions for Students** Same as above. Microdata access is not granted to students following an undergraduate program of study as opposed to postgraduate students, for whom access to microdata is given under the same conditions specified above. Legal Framework Malta Statistics Authority Act 2000: Data Protection Act 2000; Census Act 1948	NSO	**Malta**
General Conditions Conditions vary according to microdata type. While heavily anonymised microdata (public use files) are available from the CBS website, access to microdata that might allow for indirect identification of the statistical units is restricted. CBS considers all such data as "confidential" and makes them available either as scientific use files or as secure use files. The former are made for the Dutch National Data Archive (DANS), and can be delivered only to researchers in universities. The latter are accessible to legitimate researchers according to a twofold definition. First, there must be affiliation to a national institution belonging to a category specified in the law: universities, statutory organisations or institutes for scientific research, planning agencies. Institutions outside these categories can be authorised by the Central Commissions for Statistics (CCS) if they have an independent legal personality, conduct research as their primary aim, and publish results. Institutional accreditation normally lasts five years. Second, the researchers must obtain approval for the specific project they wish to undertake with CBS data. **Conditions for Non-Resident Researchers** Non-resident researchers can have access to microdata if their institutions obtain authorisation from CCS. While the criteria are the same as for national researchers, a new CCS authorisation is needed for each project. **Conditions for Students** Because access can be permitted only to researchers who are employed by a legitimate research institute, to obtain access, student needs to have a so-called "0-contract". This is because only persons with an employment contract can be held responsible by the institute they are working for. Legal Framework Act of 20 November 2003, last amended by the Act of 15 December 2004, governing the Central Bureau of Statistics (Statistics Act on Statistics Netherlands, 2003).	CBS	**Netherlands**
General Conditions Data users must be affiliated to an approved research institution or to institutions that meet a certain number of equivalent conditions. In all other cases, users must obtain accreditation for their institution before applying for data access. **Conditions for Non-Resident Researchers** Same as national researchers for fully anonymised data; de-identified data can be supplied to foreign researchers if use takes place via the country's national statistics agency, and if their confidentiality regulations correspond to those in Norway. **Conditions for Students Same** as confirmed researchers, though supervision by a qualified researcher/tutor is often required. Legal Framework Statistics Act of Norway, Act No. 54 of June 16 1989; Personal Data Act of 14 April 2000	SSB	**Norway**
General Conditions Researchers affiliated to recognised universities or other higher	Central	**Poland**

education and research institutions may have access to anonymised microdata from social surveys (no business data) for scientific and statistical purposes. The individual researcher and the institution take joint responsibility for use of the data. **Conditions for Non-Resident Researchers** Non-resident researchers from the EU and affiliated countries may apply for access under the same conditions as national researchers. Contract procedures and requirements are slightly simplified (in terms of signatures, in particular), but the information to be provided, obligations, and fees are the same. **Conditions for Students PhD** students and higher may apply for access to data. Legal Framework The statistical law of Poland (1995) emphasises protection of confidentiality, especially of businesses. The 2014 CSO regulation offers a possibility of access for scientific research purposes.	Statistical Office of Poland	
General Conditions To facilitate researchers' access to data, INE has a protocol with the Ministry of Science, Technology and Higher Education, specifically the Foundation of Science and Technology (FCT - entity responsible for funding R&D in Portugal) and General Directorate of Statistics for Education and Science (DGEEC). The protocol concerns researchers from universities and other legally recognised higher education and research institutions. DGEEC is responsible for accrediting users and providing them with the necessary information. The researchers must sign a form (online submission will be available soon) and a Statement of Commitment (each researcher involved in the request must sign one). The accreditation granted by DGEEC is valid during the declared length of the research project and only for the data identified in the request. It requires signature of a Code of Conduct by the applicant and the research institution of affiliation. Under the protocol four access modes are authorised including provision of fully anonymised data files and ready-made tables that allow no form of re- identification of statistical units; access via a secure remote access IT system to data that enable accredited researchers to build customised tables; and exceptionally, onsite access in a safe environment, allowing use of indirectly identifiable microdata under strictly controlled conditions (subject to a previous additional assessment by Statistics Portugal and an external group of experts in the area of the request). **Conditions for Non-Resident Researchers** Non-resident researchers can access statistical data under the same conditions as the Portuguese if they are in a Foundation for Science and Technology Portuguese training scholarship or if they participate in co-operation programmes in R&D with Portugal. **Conditions for Students** PhD students have access under the same conditions as other researchers. Master's students need to fulfil an additional condition: the request and the statement of commitment must be also signed by the supervisor Legal Framework Law of the National Statistical System, NSS (Law No. 22/2008 of 13 May, article 6o, no 7 and 8)	National Institute of Statistics	**Portugal**
General Conditions Microdata can be accessed for research purposes. Researchers have to be affiliated to a research institute, university, National Statistical Institute, Central bank, consortium that has a partnership with an accredited research institute for a specific research project. Research institutes are accredited by the National Authority for Scientific Research (ANCS) which is under co-ordination of Ministry of Education and Research. **Conditions for Non-Resident Researchers** Same conditions as national researchers. A Statistical Confidentiality Committee (SCC) considers requests that are not covered by legislation (SCC has a consultative role). C**onditions for Students** PhD students who are conducting research projects co-ordinated by scientific researchers may have access to data. Legal Framework National: Law no.226/2009 for organisation and functioning of official statistics in Romania; Law no.677/2001 for the protection of people, processing personal data and the free circulation of these data with the subsequent changes Accreditation conditions G.O. 57/2002 and G.D. no.551/2007	INSSE	**Romania**
General Conditions Research institutions, universities and other higher education organisations may have access to confidential statistical microdata for scientific purposes. Members should request the data via the research institution that employs them. SR may provide access either by sending anonymised versions of the data to	SOSR	**Slovakia**

the requesting institutions (on CD-Rom or USB device) or by allowing access to de-identified data removed at its onsite safe centre. **Conditions for Non-Resident Researchers** Non-resident researchers may access data under the same conditions as national researchers. **Conditions for Students** PhD students may have access under the same conditions as confirmed researchers. Legal Framework Law No. 540/2001, as amended (art. 30).		
General Conditions Microdata can be provided for research purposes to registered research institutions and registered researchers in both academia and government offices. "Registered" researchers have a national identifier. As a general rule, users submit applications for access that are assessed by a Confidentiality committee (internal to SORS); the latter prepares recommendations for SORS's board of directors, which makes the final decision. **Conditions for Non-Resident Researchers Same** as national researchers (though requests from non-registered researchers should be evaluated by the internal Confidentiality committee). **Conditions for Students** Same as confirmed researchers, provided a tutor/supervisor signs the access agreement on behalf of the student. Legal Framework National Statistics Act (OJ RS, No. 45/95 and 9/2001)	SORS	**Slovenia**
General Conditions Access to data containing direct identifiers is never granted. Access to data allowing for indirect identification of statistical units may be granted, under some conditions, to researchers employed in a research capacity by an eligible University or other Higher Education or Research Institution. **Conditions for Non-Resident Researchers** Access is allowed only when a recognised research institution signs the request or when an official regulation exists. **Conditions for Students Access** is allowed only when a recognised research institution signs the request or when an official regulation exists. Legal Framework Spanish Statistical Act 1989. Spanish Act on Protection of Personal Data 1999 Spanish Regulation on Protection of Personal Data 2007	INE	**Spain**
General Conditions Access is provided to researchers affiliated to recognised higher education or research institutions in Sweden. **Conditions for Non-Resident Researchers** Access is restricted to certain data files for EU researchers. No access for non-EU researchers. **Conditions for Students** No access for students Legal Framework Secrecy Act § 19 Data Protection Act	SCB	**Sweden**
General Conditions Data can be supplied for research purposes only. **Conditions for Non-Resident Researchers** Non-resident researchers must additionally demonstrate security measures planned for data protection. **Conditions for Students** Same as for national researchers. The student's supervisor (Professor) also signs the contract with SFSO. Students are entitled to a 20% discount. (For students registered in a foreign university, it is recommended that the supervisor acts as principal investigator). Legal Framework Federal Statistics Act of 9 October 1992, n. 431.01	BFS	**Switzerland**
General Conditions Data can be accessed for research, teaching and analysis purposes. **Conditions for Non-Resident Researchers** Same as for national researchers. **Conditions for Students**: Same as for confirmed researchers, if working under the supervision of a tutor.	ONS	**United Kingdom**

6.4.4 ACTORS IMPACTING COVERAGE

Many elements can impact the level of coverage presently accomplished by CESSDA and CESSDA's SPs amongst which the development of proper co-operation agreements between

NSIs and local SPs, as it was stressed in past projects. However, other elements impact how well CESSDA and CESSDA's SPs are covering the field.

The emergence of various actors, the use of new devices by researchers, or yet again the relationship between administrative services and local infrastructures providing DASs are a few of many examples that stress the evolutions affecting coverage provided by CESSDA. For instance, SSH scientists used to be mainly employed in the public sector and would privilege the use of DAS in this sector. Some countries are better allocated in funding in Europe than others i.e. UK and in the northern European countries (Sotiropoulos, 2006), which helped explained the massive data production stemming from these countries, pouring over to their related national archives. But with technological changes and private corporations coming heavily into the data production fields, the landscape of the market of information is evolving at a rapid pace[75]. Private corporations are not subjected to the rules set out for the public sector. Instead, they follow the rules of the market. As the report of the Social Sciences and Humanities for Europe puts it, "*dissemination and distribution functions have an ad hoc character*"[76].

6.4.5 NATIONAL LEVEL POLICIES & RELATED STRATEGIES

Legal conditions constitute an essential precondition for research access to take place and must therefore be considered albeit only briefly. In the case of official statistical, it appears that all countries investigated in 2015 during the DwB (i.e. EU, EEA, EFTA members) offer research access. This access, built off national legal frameworks, guarantees that researchers, pertaining to certain conditions, can use OS for scientific purposes[77].

Law on statistics - usual contents[78]

It was established during DwB that:

"*European laws and regulations, as well as the statistical laws of individual member states, recognise the data needs of research and include provisions to offer (regulated) access to official microdata, including confidential data, for research purposes. Typically, these data would not be accessible to anyone outside the NSI that collected and processed them: in this sense, research purposes constitute an exception.*"

It is further mentioned:

"Most statistical legislations require research purposes to be non-commercial, some of them demanding evidence of that such as institutional mission of applicant's employer organisation or publication of outputs. Cognate activities that may or may not be recognised as research, with

[75] Report (2008) of the Project: Social Sciences and Humanities for Europe (SSH futures) Instrument: Specific Targeted Research Project, Thematic Priority: 7 Citizens and Governance in a Knowledge-Based Society, http://www.iccr-international.org/sshfutures/docs/SSH-FUTURES
[76] *Ibid.*
[77] DwB D3.3 Researcher accreditation - current practice, essential features, and a future standard p. 20.
[78] http://www.dwbproject.org/export/sites/default/edaf0/dwb_edaf_session-h_access_across_borders.pdf
Tomaz Smrekar, "*Access to statistical micro data across borders: an introduction about the legal issues*", 1st European Data Access Forum, 28 March 2012.

differences across countries, include public policy evaluation as well as teaching and learning. Please check the individual country factsheets for details"[79].

Fundamental principles of national statistics can be summarised in the following manner:

- neutrality, objectivity, professional independence, rationality, statistical confidentiality, transparency, etc.,
- organisation and status of national statistics - authorised producers of national statistics, their functions (collecting and processing of data, dissemination of statistics), Head, Statistical Council,
- Funds and expenses of the national statistical system,
- Programme of statistical surveys,
- Data providers (including holders of administrative records) - reporting duties, rights,
- Storing of data (including personal data),
- Information security requirements,
- International statistical co-operation,
- Penalty provisions (for non-providing data, for misuse of data).

Law on statistics: microdata access for research purposes

- Sometimes the law is silent (old legislation, countries with limited demand for micro data use),
- Many times, mentioned without details,
- Sometimes even organisational details are described. Law on statistics: micro data access across borders for research purposes[80].

6.5 Examples of co-operation between Data Archives Services and Official Statistics

In this section, we provide some good examples of a progress in co-operation among data archives services (DAS) and official statistics (OS) focusing mainly on Nordic countries. Progress in this field is also present in the cases of Hungary in terms of the legal framework regulating access to anonymised personal data, as well as in the case of the Czech Republic in terms of the co-operation between DAS and OS.

6.5.1 Nordic Countries

In the Nordic countries can observe recent notable efforts towards establishing common (Nordic) microdata-related services, based on common metadata standards and tools. organisational models and agreements reached between DAS and OS, supported by changes in legislation on a country level, lead to increased (cross-) country micro-data access, in particular to census and administrative (register) data.

[79] http://www.dwbproject.org/export/sites/default/service/accreditation_db_pdf/dwb_accreditation-factsheets_glossary_july2014.pdf

[80] http://www.dwbproject.org/export/sites/default/edaf0/dwb_edaf_session-h_access_across_borders.pdf Tomaz Smrekar, "*Access to statistical micro data across borders: an introduction about the legal issues*", 1st European Data Access Forum, 28 March 2012.

6.5.1.1 CHANGES IN LEGISLATION AND COOPERATION BETWEEN DAS & OS

There were some legislation and organisational changes over the past few years. Finland reported on Statistics Act of 2004[81] that entered into force in September 2013. Changes affected the possibilities to use data from Statistics Finland for academic research purposes, mentioning both release of confidential data for scientific research, and production and release of files intended for public use. Statistics Finland launched a remote access system in 2010 and after Archival Act in 2013, researchers switched to using data over the remote access system, which enables them to use data where only direct identifiers have been deleted. Statistics Finland is working on developing the system; numbers of users and the size of data sets are steadily growing.[82]

Statistics Norway (SSB) together with The Norwegian Social Science Data Service (NSD) have started a project, called RAIRD – Remote Access Data infrastructure for research Data which will influence on organisational changes.[83] Since 1975, Statistics Norway and NSD collaborated formally in facilitating and distributing SSB's data to Norwegian research institutions, thus providing researchers with extremely good conditions compared to those of colleagues in other countries[84]. As example of good practice, NSD's agreement with Statistics Norway on the dissemination of data for research purposes can be exposed. NSD emphasises close collaboration with several Norwegian bodies and organisations. This collaboration is unique in the international context[85] and will be described in the following chapter more closely.

There were also organisational changes in 2014 in Denmark. Danish Data Archive is now responsible for all digital born data in the National Archive, both survey data and official statistics. Danish Data Archive was given key role regarding curation and access services of the official microdata, including data collected at Statistics Denmark. According to the Archival Act, Statistics Denmark and The Health Data Centre, have to deposit their data to DDA for a long-term preservation.[86]

6.5.1.2. MODES OF COOPERATION WITH OFFICIAL STATISTICAL BODIES

"The Norwegian Centre for Research Data (NSD) cooperates closely with several national public institutions to provide research and educational communities with comprehensive high-quality data products and services. The Research Council of Norway support NSD financially as the most significant national infrastructure facility for research data archieving and access. NSD holds large amounts of data on individuals, regional units and also the political and administrative system of the country, covering more than 200 years of the history. NSD cooperates with Statistics Norway as their chosen channel for dissemination of data for purposes. Likewise, NSD also holds a formal role as a facilitator and trusted archive between research and the Data Inspectorate, resulting in extensive collection, use and archiving of microdata for research purposes.

[81] http://tilastokeskus.fi/meta/lait/2013_tilastolaki_en.pdf?_ga=1.25796735.394599202.1476259379
[82] DwB survey from Statistics Finland
[83] DwB survey with Statistics Norway
[84] http://www.nsd.uib.no/om/doc/nsd_annual_report_2015.pdf
[85] http://www.nsd.uib.no/nsd/doc/nsd_annualreport2013.pdf
[86] Correspondence (e-mail) from Christian Lindgaard Olesen – Archival Act (https://www.retsinformation.dk/forms/r0710.aspx?id=12066)

Similarly, Danish Data Archive (DDA) is responsible for dissemination of all digital data in the National Archive (both survey data and official statistics). DDA also collects datasets from Statistics Denmark (estimated to approximately 40 datasets per year) and the National health registers, as soon as they are published[87]. Data collections, distributed by DDA, thus include all three major official statistics data providers in Denmark.

On the other hand, the number of official statistics surveys in the distribution of Swedish National Data Service (SND) is rather scarce[88], however, SND co-operates with statistics producing bodies, aiming to gain a comprehensive overview of the databases available[89]. Finnish Social Science Data Archive (FSD) is currently not yet distributing official statistics microdata, although they do have some data from statistical office and organisations that produce official statistics[90]. Yet, Finnish data archive is part of Statistics Finland steering group, working towards establishing microdata remote access system[91].

Building a national research infrastructure for access to OS microdata is one of the main goals of Norwegian archives as well. NSD and Statistics Norway are working on a project Remote Access Data Infrastructure for Research Data (RAIRD)[92]. Project aims to provide easy (remote) access to large amounts of rich high-quality statistical data for scientific research, giving researchers possibility to analysis microdata with accompanying metadata, while at the same time managing statistical confidentiality and protecting the integrity of the data subjects[93].

6.5.1.3 ACCESS TO MICRODATA (WITH FOCUS ON REGISTER BASED MICRODATA)

DDA Search[94] provides largest collection of survey-based research data for researchers and students in Denmark. DDA also provides access to register based data through National Archives Database Daisy[95]. Users have open access to archival materials, earliest 20 years after the last dated record in the dataset. In case the materials contain sensitive information (e.g. income, health or other personal conditions), waiting period for open access prolongs to 75 years. However, researchers are eligible to apply for access to datasets that are not yet open to public. The Danish Data Protection Agency issues the permission, and DDA can then decide on distributing the original datasets, while researchers obliged themselves to protect individual information before publishing research results[96]. DDA is also working on online ordering system for micro data[97].

In Sweden microdata are protected by the Security Act and researchers should apply for access through Microdata Online Access (MONA)[98] - standard system for documentation of

[87] E-mail correspondence with Christian Lindgaard Olesen
[88] E-mail correspondence with Iris Alfredsson
[89] https://snd.gu.se/en
[90] E-mail correspodence with Helena Laaksonen
[91] DwB survey - FSD
[92] DwB survey - SSB
[93] http://raird.no/whitepaper/whitepaper-detailed.html
[94] http://dda.dk/simple-search?lang=en
[95] https://www.sa.dk/daisy/daisy_forside
[96] E-mail correspondence with Christian Lindgaard Olesen
[97] DwB survey – DDA (check)
[98] http://www.scb.se/en_/Services/Guidance-for-researchers-and-universities/MONA/

microdata, operated by Statistics Sweden. SND enables access to microdata in their own distribution (with various restrictions applied) as well as provides access to metadata on surveys that are not distributed by SND but through other data providers (both institutional and private, such as researchers themselves)[99]. Currently, SND is working with the Swedish Research Council (Vetenskapsrådet) on building a service for researchers to get access to data from registers[100]. Registerforsning.se is information portal for researchers who want to use Swedish register data in their research. The portal provides information on registers and their contents, guidance on how to access data, lists of register keepers, a description of register research as well as information on current legislation[101]. A dedicated tool, that will allow researchers to get detailed information on Swedish registers on metadata level, is promised to be available via the website during 2016.

Aila[102] provides access to datasets archived at the Finnish Social Science Data Archive and their study descriptions. All users can browse and search data, access study descriptions and download open access data. By registering, users are able to download data with access restrictions. Services are free of charge. Students and staff from Finnish universities can register with Aila, using the username and password, issued by their institution (HAKA authentication). Other users apply for a username from FSD User Services.

FSD is also involved in project that aims to establish a new research infrastructure utilizing register-based data. Project was funded by Academy of Finland and involved Finnish National Archives and Statistical office. Finnish Microdata Access Services (FMAS) is a new research infrastructure, intended to facilitate access to register data and increase confidentiality protection in research utilizing register-based data. In 2013, FMAS was accepted into Finland's roadmap for research infrastructure. The project is hosted by the National Archives, as well as Statistics Finland, and is funded by the Academy of Finland. FMAS will greatly increase the number of high-quality, register-based studies, and the potential for new innovations that will increase the competitiveness of Finnish science. The infrastructure will also improve prerequisites for evidence-based policy.[103]

NSD - The Norwegian Centre for Research Data (NSD) cooperates closely with several national public institutions to provide research and educational communities with comprehensive high-quality data products and services. The Research Council of Norway support NSD financially as the most significant national infrastructure facility for research data archiving and access. NSD holds large amounts of data on individuals, regional units and also the political and administrative system of the country, covering more than 200 years of the history. NSD cooperates with Statistics Norway as their chosen channel for dissemination of data for research purposes. Likewise, NSD also holds a formal role as a facilitator and trusted archive between research and the Data Inspectorate, resulting in extensive collection, use and archiving of microdata for research purposes.

[99] https://snd.gu.se/en/deposit-data/accessibility-levels
[100] http://www.registerforskning.se
[101] DwB survey - SND
[102] https://services.fsd.uta.fi/index?lang=en
[103] http://www.arkisto.fi/en/finnish-microdata-access-services

While register based data available in Sweden[104] and Finland is currently largely complied for Nordic Health Data project, NSD provides access to microdata, collected from registers, e.g. welfare data, such as data from the National Insurance System, the Social Welfare System, and the Labour Market register, data from the Regular GP Scheme, Data from the Patient Ombudsman System, NSD'S Generation Database, NSD's Census Data Bank[105].

"Making Nordic Health Data Visible" is a cross-country collaborative project, which will make easier to identify and locate health data in the Nordic countries. Half-way through the project, a pilot web service that permits metadata to be harvested from a number of sources has already been completed. This will make it possible to search for and find health data in the archives of all Nordic countries. The catalogue service has been designed in such a way, that it will also collect information from sources outside the Nordic region, and even from other fields than health. This approach was discussed in the DwB project, to be extended on pan-European official statistics data sources. Emphasis has been laid on making it as user friendly as possible via a very simple user interface, and by offering pre-defined keywords and concepts, that are particularly relevant to health data. The two-year project, which will end in September 2016, is led by NSD and is financed by NordForsk[106].

6.5.1.4 METADATA

Although metadata standards are not unified among different institutions (statistical offices, data archives and other data collectors), we can observe several attempts in strengthening co-operation among stakeholders, aiming to establish better and unified metadata standards.

Statistics Denmark and Danish Data Archive (DDA) provide a good example of effective co-operation. In 2015 Statistics Denmark reported on a change in progress, by starting to build a metadata portal. Both organisations are heading to common metadata standards, which is, according to DDA perspective, a must. In 2011 DDA initiated project, aiming at a common integrated metadata system, in order to improve quality and facilitate dissemination of statistics. At the beginning of 2015, Statistics Denmark launched a DDI-based system handling concepts and quality information for 237 statistics.

The roadmap towards common metadata in a statistical context requires, as discussed, both improvement and precision in the terminology when talking about metadata, and better understanding of role of metadata in relation to users.[107]

Finland, on the other hand, report difficulties regarding metadata for register based data, because the registers are scattered across many institutions[108]. In spite of this, there has been intensive co-operation between Finnish Data Archives and Statistics Finland, especially with regards to metadata and with regards to availability of data gathered by Statistics Finland through Data Archives.[109] Both Statistics Finland and FSD produce good-quality metadata for their research data. However, the metadata and the systems are not directly compatible, as

[104] Correspondence with Martin Brandhagen (check)
[105] http://www.nsd.uib.no/nsd/english/individualdata.html
[106] http://www.nsd.uib.no/om/doc/nsd_annual_report_2015.pdf
[107] Mogens Grosen Nielsen, Flemming Dannevang, 2015-11-20 – e-mail conversation
[108] E-mail from Annaleena Okuloff 9.6.2016
[109] DwB survey (Statistics Finland)

both organisations apply their own metadata practices and models. In other words, metadata produced by FSD is not as such usable by Statistics Finland, and vice versa. The missing interoperability complicates collaboration efforts, such as having a common data catalogue. Future goal is to find solution for better, more comparable and easier collaboration.[110]

In 2010 Statistics Sweden decided on a register and data warehousing strategy with the focus on activities, directed towards structured metadata of good quality, that will enable to identify relevant variables, populations and object in the data store.[111]

6.5.1.5 PUFs & SUFs

Few information was reported about PUFs and SUFs. Statistics Finland can produce and release to public use such files, formed from data collected for statistical purposes, from which identification data have been removed and which have been processed so that the statistical unit cannot be directly or indirectly identified.[112] However, FSD is currently not disseminating official statistics data, and it is not certain whether some of the data used for producing official indicators is accessible to user. According to FSD, there has also been discussion about co-operation in this area in the future with the SUFs, while PUFs are not produced, as they find their value questionable.[113]

Danish data Archive (DDA) does not operate with PUFs or SUFs. However, DDA does have an obligation to make their data available to scientists. If they receive permission to get the data, they get the original version of the data (with sensitive information). DDA preserve the data in their own version of SIARD format[114] and they guarantee that data can be preserved indefinitely. DDA is still in the process of creating procedures for converting the data to statistical software packages, as well as anonymising it, as a service to scientists[115].

6.5.2. HUNGARY

In Hungary, the law ensuring access to anonymised personal data for research and policy analysis, introduced in 2007 (Scharle, 2017), could be considered as an example of good practice, enabling researchers and policy makers that their interests are recognised and acknowledged. However, the law implementation process had to overcome several obstacles that were primarily rooted in the lack of trust between academic and government organisations, as well as in the strict legislation on personal data protection. Despite significant amounts of administrative data being collected on regular basis, the use of the data was scarce. Following the law that omitted the use of data for research or statistical analysis, data owners were not able to anonymise personal data on legal basis. First initiatives to change legislation arose when the Finance Ministry expressed its interest in evidence based policy making and liaised with various stakeholders: data owners, potential data users (analysts in the civil service and researchers), the Ombudsman for data protection and the National Development Agency among others. The idea of providing access to micro data had to face strong opposition in

[110] http://www.fsd.uta.fi/fi/julkaisut/julkaisusarja/FSDjs11_metadata.pdf
[111] E-mail correspondence with Iris Alfredsson, wrote on 10.6.2016
[112] DwB survey (Statistics Finland)
[113] Correspondence via e-mail with Helena Laaksonen
[114] http://www.digitalpreservation.gov/formats/fdd/fdd000426.shtml
[115] Correspondence via e-mail with Christian Lindgaard Olesen, 9.6.2016

several other stakeholders: Ministry of Justice, Ministry of Education, the Central Statistical Office, the State Reform Committee and the Ministry of Economy, to name some of them. After approximately a year of intensive negotiation process, the new law passed the Parliament, however, with the Central Statistical Office remaining sceptical, naming the reliability of the anonymisation as the main concern to oppose. The new law now enables data owners to process personal data for the purposes of anonymisation and along these lines supports evidence based policy making and research with taking into consideration personal data protection.

6.5.3. CZECH REPUBLIC

In Czech Republic, CSDA has established co-operation with the Czech Statistical Office (CZSO) aimed at linking data services between these two institutions. The Czech Statistical Office (CZSO) is a central body of the Czech Republic state administration. Currently CSDA has been publishing detailed information about selected scientific datasets available at the CZSO. On the CSDA website there is already information on the CZSO data available for scientific and academic purposes, along with survey questionnaires. Soon, metadata from the selected statistical research will be added to the Nesstar catalogue. In the more distant future, the CSDA catalogues are supposed to be used to search for the CZSO data, including primary data. The framework agreement concerning further co-operation and publication of the CZSO metadata in the CSDA system – Nesstar, is being under preparation. The proposed agreement on co-operation between CSDA and CZSO bind the parties, among others, to:

"CZSO will hand over metadata from statistics provided by the CZSO. Moreover, CZSO will allow for publishing those metadata by the CSDA data archive. CSDA is a part of the Institute of Sociology, Czech Academy of Sciences, and serves as a data centre for scientific research. Data publication will be possible through the international consortium CESSDA (CSDA is an active member of the consortium), as well as in related international projects. CSDA will: publish and promote information on the CZSO services, providing with statistics for scientific research; publish the CZSO metadata for the CSDA publishing; attempt to make these materials more visible for further, secondary use of databases in the field of scientific research, higher education and scientific training. CZSO will provide with professional consulting and advice support, when it comes to data search from social research projects and its usage concerning the activities provided by the CZSO".

6.6 CONCLUSIONS

As can be seen from the report, there are to be find examples co-operation among data archives services (DAS) and official statistics (OS) in all field that have a potential for co-operation. Some of activities extend beyond individual countries. Project with the aim to provide solution for a search portal, access to administrative register microdata, and co-operation on metadata production can be pointed as outstanding.

7. HISTORICAL DATA

7.1 SUB-TASK DESCRIPTION

The objective of the task is to support the data services to widen their data perimeter. The task reviews the state of play regarding the diversity and the amount of historical data covered and it identifies the obstacles encountered regarding them.

7.1.1 DESCRIPTION OF HISTORICAL DATA

Historians' research interests change over time, and there has been a shift away from traditional diplomatic, economic, and political history toward newer approaches embracing social and oral history aspects. History as a discipline can be related to humanities as well as to social sciences thus in that sense, it is situated in the border of both scientific domains. This approach reflects CESSDA Archives understandings as well, as we will explore further in this section of MS9 (see section 6.2).

A primary target is to trace the frontiers and intersections between social science and historical data to better understand the boundaries of historical data nowadays. Then, we will explore what forms of co-operation should be set up between CESSDA and the major actors in this field.

7.1.2 METHODOLOGY: FROM REVIEW OF THE LITERATURE AND INTERVIEWS TO EXPLORING CESSDA ARCHIVES

We followed a threefold methodology in order to explore the current state of play regarding historical data. Firstly, the two partner organisations involved in task 3.4 to handle this domain, i.e. EKKE and ICS, carried out a small scale qualitative research. Namely interviews with historians were conducted in order to initially investigate and map the field of historical studies. EKKE research team has carried out two interviews; the first one is an academic staff, full professor in a major Greek university regarding social sciences, while the second one is a researcher in the field of history at EKKE. ICS' research team has also conducted two interviews with researchers in the field of history from ICS. The results of the interviews are presented along with data analysis in the respective sub-chapters.

Secondly, we navigated through the web the CESSDA Archives, in order to:

a) Locate historical data.
b) Identify, if possible, the proportion of historical data in relation with the total number of datasets in each archive.
c) Identify the main producers of historical data.
d) Identify other actors that produce historical data.
e) Highlight national policies and related strategies regarding data dissemination.

Thirdly, we proceeded to a literature review in order to further explore data policies at national and international level, locate historical datasets in other archives and discuss the current trends.

7.2 SCOPE OF DATA DOMAIN: MAIN UNDERSTANDINGS & NEW ISSUES ARISING

7.2.1 APPROACHING HISTORICAL DATA

History is a peripheral discipline for CESSDA to the extent that it belongs to the humanities rather than the social sciences, which constitute member Archives' main business. However, there are intersections between historical and social science research that require scholars to access data from both recent and past sources. What's more, some CESSDA Archives have historical data in their holdings and have experience and expertise in this area (Kondyli et al, 2012).

The definition of frontiers between social sciences and historical data presents to be a difficult task, as there is no specific/ systematic guidance for classifying a dataset as historical. In other words, what seems to be an important issue, at least for CESSDA Service Providers, relates to classification and/or terminology. When searching historical datasets, with the exception of few SPs, the way they classify relevant datasets lack clear terminology. It seems important to get again the classification thesaurus (ELSST) in full operation in order for the SPS to refer to the same concepts as well as to facilitate users to their search. Thus, we speak about quality and efficiency in providing data services as a whole.

The distinction is essentially based on methodological issues and the scope/focus of the dataset. Social science data have usually a broader focus and they seek to approach and/or interpret the social world (by exploring for example attitudes, perceptions, practices, etc.) by applying both quantitative and qualitative research tools, such as surveys, interviews, content or discourse analysis. Historical research, as it is stated in Gesis (Franzmann, 2015), seeks to systematically describe and analyse past societies through the usage of theories, formal methods and quantitative data. The data is generally not obtained through surveys but by census and processing historical sources. According to Franzmann (2015), historical studies differ in the following aspects from survey studies.

a) Investigation period: In survey research, the investigation period is the time period during which survey was carried out, and therefore, it is the time period in which the data was collected. In the case of historical studies, the period of investigation refers to a time period lying in the past, for which retrospective data are collected from official or non-official sources.

b) Universe and Selection Method**:** Data of historical studies are not collected by interviewing persons – with the exception of oral history, - they are collected from archives, publications of the official statistics and/or other statistical material and sources. Thus, the source types are for example archival documents, record collections, church registers, official statistics, scientific publications, etc.

According to one of the researchers interviewed, we can say that social sciences and historical data intersect at two levels:

1) Methodological level: Differences in research methods between social sciences and history are increasingly smaller. For example, quantitative methods of statistical and text analysis are more and more used by historians in their research.

2) Thematic level: Research subjects that are traditionally at the centre of interest for social sciences, such as social inequalities and exclusion are now an interdisciplinary common ground for economic historians for example and social scientists.

To sum up, while boundaries were previously rather distinct in regards with methodological treatments and thematic approaches, distinctions between data fields have progressively diminished. Still two main features remain regarding historical data (1) period of investigation and (2) documentation as primary source.

7.2.2 TRENDS AND STAKES IN RESEARCH: TIME SERIES, AGENT-BASED APPROACHES AND ICT OPPORTUNITIES

In the field of history, archives form a major component of research work. Hence, humanities research and the use of archived materials tend to go hand in hand (Fält, 2015). Archives are at the very beginning of the data lifecycle to researchers by providing them primary sources such as original manuscript documentation. Naturally, the outcomes of such research can vary. Hereinafter, we examine some of the new trends in research to better foresee the use of datasets and start under covering researchers' needs.

The development of time series: a trending use of historical data

According to our key informants, in the last few years different research agendas in history have been converging to the understanding that production and dissemination of "time series" are important within the discipline. These series are built on quantitative data – gathered from diverse historical sources, documents and archives, such as prices, wages, products, patrimony, wealth and employment. The subject of these series can range from agriculture to economy or politics. Economic historians are those who are more likely interested in this kind of data than historians of ideas and thought.

As an example of a research initiative of this kind, ICS is involved in a research project that is being co-ordinated by an Australian university in partnership with the Australian NSI. The project intends to collect data and build transnational datasets concerning wine consumptions, production and export trade across time. Another example is the "Price, Wages and Rents in Portugal 1300-1910"[116].

Many factors have contributed to make the use of time series a trend in the discipline. On the one hand, stabilization of standardising procedures of historical information in order to measure issues such as consumption, income, investment and saving, has led to use time series more and more. On the other hand, the consensus regarding the importance of quantitative data and specifically time series to (i) systematise historical information, (ii) promote the continuity of research and (iii) understand contemporary trends.

[116] http://pwr-portugal.ics.ul.pt/?page_id=56

New approaches applied to old datasets: building categories of historical data

In another perspective, Tanaka (2015) raises the issue of how things change and how historians describe that change.

By examining a database of recorded happenings, he explores a period of time in Japanese history (1884) that it is not a traditional historic landmark. He explores the very formation of the categories "old" and "new" and set a different way to think about change beyond the linear and celebratory narratives of Western systems. Various happenings, now called as folklore and superstition, were very much of the present. But along with this world where the past and present were indistinguishable, others identified these same objects, things and ideas to be old-fashioned. Simultaneously, the old was separated into the dead past and heritage, aiming at formulating a historical past for the emerging nation state. Thus, 1884 was empty of 'important" events in terms of the political and economic development of the new nation-state. Each happening embeds meaning within a particular place or moment, showing isolated or overlapping temporalities. The past in this case is a layering of different temporalities: that in which place and immediacy is preeminent, in contrast to our current practice of incorporating objects and people into a singular narrative of national becoming. Here, the database raises questions about standard historical narratives. The fragmentation of inherited knowledge into different dead pasts, heritage and tradition, the relegation of ghost stories to the category of superstition and later the field of folklore, all served the political purpose to build a strong nation-state by educating (civilizing) the masses.

Tanaka argues that history downplays individual culture and he quotes Simmel *"The things that determine and surround our lives, such as tools, means of transport, products of science, technology and art, are extremely refined. Yet individual culture, at least in the higher strata, has not progressed at all to the same extent; indeed, it has even frequently declined"*. Thus, Tanaka concludes that despite the advances to bring common people into history, history is still oriented towards describing the tools, transport, science, technology and art of the newly formed nation-states.

Agent-based focuses

Another way to approach historical evolution and change focuses on agents. However, in order to understand the role of agents, we have to recognise the situatedness of information the meaning of which, when extracted and placed in a different setting (the creation of the modern archive and the historical "fact"), is often altered. This shift affects the type of data selected for research.

The use of ICT: a new actor changing the playing field

Historical data field is also affected by the advance of technology and the development of archives. These changes offer researchers new possibilities for recovering stories and storytelling in history and grasping in a genuine way the sensibilities of another era, e.g. how agents confronted with something new and unexpected. Such sources of information, as Tanaka notes, could be for example personal stories, diaries, writings or records that "embed" facts, or, in other words, give us alternative pathways for approaching the traditional historical facts (2015: 27).

7.2.3 MAIN SOURCES OF DATA OF INTEREST: RESEARCH SOURCES AND METADATA

With regard to data, the interviews highlighted two main needs:

a) Quantitative research data in history and recommendations: quantitative research data roadmaps, open access and archiving time series

Concerning the first case, one of our key informants mentioned how important would be the establishment of a road map for sources in history in order to meet projects and institutions involved in the field of historical data, namely those who produce time series. This is seen as relevant both at the national and international level. According to the same interview, time series are the research outputs that have more interest in historians and, at the same time, are more likely to be used by other researchers from the social sciences. In this sense, they are the research outputs that seem particularly important and worthy to be added to CESSDA collections.

Bearing in mind all the expertise developed by CESSDA for social sciences, this trend could be an opportunity for CESSDA to invest in new analysis tools as GESIS provides with Histat – Historical Statistics. In brief, histat integrates numerous historical studies with time series data into a database, which is subject and study orientated. Database contains a wide selection of topic titles and individual studies are allocated to these topics according to their thematic subject (Franzmann 2015).

In the case of another key informant, the idea of a central repository (open data) would be a best practice example though historians seem to be less familiarised with more innovative tools. One of the interviewees also mentioned that National Statistics Institutes (NSI) can also be important data providers for contemporary history since they have collected information for a long time now. The extent to which researchers can access to historical data and documentation from NSI and how do historians relate and engage with such organisations is a topic that requires further investigation since it also relates with social sciences researchers needs.

b) Metadata about research data and recommendations: the role of CESSDA in sustainability matters

Metadata for historians are a matter of great importance. This is related to the fact that historians can use all sort of historical sources in their research – from public documents to private letters – since data was not collected for statistical purposes in previous historical periods; The so called "statistical thought" only became common in the mid of the nineteenth century. Therefore, metadata ensures the quality of the data for purposes of validation by identifying conversion procedures of the data used, source, place and origin of the data. Metadata and the proper archival of research data allow other researchers to return to sources that are no longer available in historical archives. In fact, one of the interviewees reported that continued access to historical sources is not granted forever by some of the historical archives. This raises sustainability issues.

It is noted that sustainability does not just mean keeping the data alive, but enabling the exploitation of advances both in technology – making the data accessible in new ways – and

forging connections between resources that lead to new discoveries and broader impact. This is essential to ensure long-term interest and sustainability of these resources (Marker, 2015). According to previous reports (Kondyli et al., 2012) CESSDA-ERIC strongly relies upon the continuous knowledge of data producers and the latest trends in research data. As pointed out in the same report, CESSDA should at least work as a hub to locate information.

7.3 From individual to national datasets: agreements and related strategies

7.3.1 Proportion of data currently archived by the existing data services

The issue addressed in this section is the type of data covered by Members countries. According to previous reports (Kondyli et al., 2012), politics is an over-represented subject in most of CESSDA Archives, while less than 50% have collections concerning History, Information and Communication, Transport, Travel and Mobility, etc. The under-representation of some topics should be put into perspective with the fact that in the same period, only seven CESSDA-members had more than 1,000 datasets in their collections. To have a better understanding of what CESSDA and Member countries' Archives contain, to later further identify what needs to be added to respective collections, we explore the proportion of data currently archived by the existing data services.

We mostly followed data sets' classifications given by researchers and/or archive services, when provided, in order to classify datasets as "historical". In some cases, we followed ELSST in order to suggest specific themes. Datasets regarding demography/census/population have been classified as historical from a time perspective, following the classification of Gesis. Archaeological datasets are excluded within the frame of this specific work.

Based on the aforementioned, we made a first attempt to map Data Archives' historical landscape. As there is usually no clear categorisation of historical data in the Archives, this attempt met with significant difficulties and will be revised in the future under the light of the information provided by the Archives. However, it still captures the large picture of historical data in the Archives.

- **ADP** provides about 600 datasets, while **CSDA** does not specify the total number of datasets offered. Both Archives do not provide historical datasets directly through their research infrastructures. FSD provides one historical dataset in a total number of about 1300 datasets, while TARKI provides a restrained number of historical datasets for a total number of about 650 datasets.

- **Gesis** (Histat), **Dans** and the **UKDA** are the major Archives regarding the provision of historical data. Histat is dedicated to the provision of historical datasets, providing a large number of time-series.

DANS provides in total a big number of datasets, concerning the following fields (last measurement in 31st October 2016):

Behavioural and educational sciences (1237 datasets), Economics and Business Administration (221 datasets), Humanities (32051 datasets), Interdisciplinary sciences (149 datasets), Law and public administration (785 datasets), Life sciences, medicine and health care (402 datasets), Science and technology (83 datasets), Social sciences (4565 datasets).

Historical datasets are estimated to a number of 3800. A large part of the historical datasets regards oral history, mostly covering the period of World War II.

UKDA provides 1,237 datasets classified under the subject "history". UKDA gives access to a various range of historical studies such as administrative, agricultural, religious, education, legal and local matters; it also has a rich collection of historical and contemporary censuses and longitudinal studies.

- **LiDA** seems to have the greatest proportion of historical datasets compared to the total number of datasets provided (61 out of the 300 datasets approximately – about 20%). They are mostly economical and financial data (agriculture, forestry, fish farming, trade, living indices, cash turnover), but there is also some demographical data.

- **FORS** presents a large proportion of historical datasets, approximately 700 out of the 10,000 datasets in total – that is about 7% of their holdings. Datasets are mostly documents and archives, police and court records, posters, etc.

- In **DDA**, there are about 70 historical datasets out of the approximately 10,000 datasets (about 0,7%). DDA distributes historical data that are mostly linked to demographic history and censuses; some parts deal with subjects such as immigration and minorities.

- In **SND**, there are about 30 historical datasets, mostly censuses and demographic databases in a total of more than 1200 studies.

- In **EKKE**, there are 15 historical datasets in a total number of 355 datasets (about 4%), mostly in the area of oral history.

- In **NSD**, a significant part of the historical data is assembled in some extensive databases, with data covering a range of topics, but mainly related to the political system.

a) Local political/administrative/demographic units, 250 years of most published statistics of demographic, political and economic character.

b) Biographic data of all members of parliament and government over 200 years.

c) The organizational history of the Norwegian state since World War II

- In **PROGEDO**, historical data is limited to ancient public statistics. There is no information about the number of records.

7.3.2 TYPE OF AGREEMENTS WITH PRODUCERS THAT IMPACT THE COVERAGE AND OTHER ACTORS

In most cases, individual researchers (mostly academics) are the main providers of historical data to the Data Archives. Some Archives have also made agreements with data producers. These data producers systematically provide the Archives with various kinds of data, including historical data in some cases. According to the information collected through the web investigation, some of these agreements are listed below:

- **DANS.** Statistics Netherlands (CBS) and DANS have an agreement for the provision of a large number of micro files with data for scientific research (at the individual and household units). DANS has also ensured the provision of oral history data from two major data producers, the Stichting Mondelinge Geschiedenis Indonesië (SMGI), Koninklijk Instituut voor Taal and the Veteranen Instituut.

- FSD. In 2015, Finnish copyright society Kopiosto and the FSD signed a licence agreement that allows FSD to archive and reuse material collected by researchers for their own research but created by others. This agreement applies to digital or digitalised newspaper and magazine material as well as photographs. The FSD pays a small charge per year for this licence.

- PROGEDO. Among other entities, Réseau Quételet co-ordinates ADISP - Archives de Données Issues de la Statistique Publique - that provides historical data to the archive.

- **UKDA.** UKDA works closely with owners and producers of the most important social and economic data sources in the UK to make sure they are made available to users in a timely manner. Indicatively, some of the high-profile organisations who regularly deposit data in UKDA: Department for Business Innovation & Skills (BIS), Centre for Longitudinal Studies (CLS), Department for the Environment, Food & Rural Affairs (Defra), Department for Transport, Department for Work and Pensions, Health and Social Care Information Centre, Home Office, NatCen Social Research, Northern Ireland Statistics and Research Agency (NISRA), Office for National Statistics (ONS), BMRB Social Research, UK Longitudinal Studies Centre (ULSC).

- NSD. Since 1975 NSD has had a formal agreement with Statistics Norway to act as a dissemination channel towards research and education. The general idea is that data from statistical production should be freely available for research through NSD, but the two institutions have to share the workload and expenses involved. The practical consequences are that NSD do the necessary curative work for longer term archiving and prepare data for general dissemination and statistical analysis. Based on an Agreement with the National Archive NSD holds a position as national archive and research dissemination service for digitalized research material. This cover both historical and contemporary material.

In general, Data Archives populate their collections on the basis of country and institutional regulations. It can also be the case that the profile and identity of the organisation is specifically targeted to certain areas, whereas other subjects are covered by other institutions (Kondyli et al., 2012). As far as other actors are concerned, in the Appendix 1, we list some archives outside

CESSDA that seem to hold historical datasets. In some cases, other kind of datasets can be also accessed via these archives.

In conclusion of this part on data coverage, a considerable amount of historical data is already being disseminated across CESSDA. Overall, the Archives provide numeric data of different types such as administrative records, census, time series and public statistics from the contemporary period. Some of the main subjects are: agriculture, demography, education, immigration. The level of documentation of the data collections can vary significantly and, with some exceptions, no software or tools are provided for data analysis in direct relationship with the portal of the data provider. Quantitative data in general is more likely to be made available and be reused by others. Time series are an example of such.

The majority of the actors in the field are historical archives which cannot be considered concurrent of CESSDA. They are in charge of providing historical sources and do not perform dissemination of research data. Future forms of co-operation between CESSDA and other actors in the field can range from local collaborations with data providers such as historical archives to agreements with NSI. DANS and UKDA are good examples of CESSDA service providers that have strategies in place. Further research about historical data should address the following topics:

- Specific metadata fields for historical data
- Tools for data analysis

7.4. Data sharing culture in History, national policies and research data infrastructures.

Researchers unkeen for sharing data and archiving

In humanities, which are at the centre of interest in this subtask, most researchers do not yet embrace the idea of having their own data archived and reused.

In May 2015, the FSD conducted a researcher survey to probe attitudes towards data management and archiving in health and medical sciences and the humanities. The e-mail questionnaire was sent to a total of 1,428 researchers. According to the authors, although the response rate was low (14 %), the survey still provided valuable insight into archiving and data-sharing practices and attitudes in these fields of science. 37 % of the respondents, were humanists, mostly historians (Järvelä, 2015), so while this study highlights general practices and views that could apply to the other fields of task 3.4, interestingly enough, it such much about historical data as well. According to the results, data collected by humanists typically remain in their own possession. Possible reuse, too, is mainly restricted to the researchers themselves (Fält, 2015).

a) data acquired with a lot of hard work are thought to be very personal,
b) Archiving was not consented to by research subjects in the collection phase,
c) poor data organization and documentation (Järvelä, 2015a).

However, results also indicated that half of humanists would be willing to open up their data for others to use. The humanities research tradition seems to be gradually approaching a kind of transition phase where demands for digitalisation, openness and data availability are increasing (Fält, 2015). Research now has created a large quantity of digital material that is often hosted in their home institutions, whether this institution is a public university, a private sector research centres such as Research and Development departments, or an agency of some kind, using a variety of approaches and technologies to store and share data collected. However, it is stated (Marker, 2015):

"This situation is quite dangerous, because without ongoing maintenance, a resource will cease to be usable at all as the technologies in which it was created become obsolete and unsupported. Even if the resources are maintained it is far from certain that they are in a state which allows them to be used with the most relevant techniques at the time of reuse. Access to legacy resources may be limited to a simple download or by browser access in a website".

Thus, Marker concludes that humanities data need to be made available via centres of expertise, which can provide the stability and reliability that is needed by the research community.

Unlike in some other disciplines, humanities publications do not yet require open access as a precondition for publication, although this situation may soon change. In Finland for example, funding agencies such as the Academy of Finland, promote open access policies (Fält, 2015). As Borg (2014) argues, the support of the Finnish government to open science will probably influence funding agencies to strengthen and specify their open access policies. Moreover, ministries are expected to start recommending universities and research organisations to establish data access policies.

International data policies for sharing research data in Social Sciences and Humanities.

The web-survey conducted by IFDO in 2013 (Kvalheim and Kvamme, 2013) describes in brief policies for sharing research data in social sciences and humanities and discusses possible challenges. In particular, the report refers to international data policies (emphasizing on open data policies), developed by OECD, EC, UNESCO and other agencies and groups, such as the ESFRI group, the Max Plank Society and BRTF- SDPA. The empirical part of the report is based on a data policy web-survey designed to be conducted in countries known to have academic infrastructures for data sharing in social sciences. 43 individuals from 32 countries completed the survey.

Types of funding policies

The report also identifies three types of policy statements from national science funders:

1. those that have explicit policies on data sharing and clear implementation of these policies,
2. those that have explicit policies but no clear implementation,
3. those that have no explicit data sharing policy statements.

Regarding the first case, such funding agencies are located in Australia, Canada and USA. Based on IFDO report and and our web-investigation, we collected the following information regarding CESSDA members:

- In the **UK**, funding bodies such as the Economic and Social Research Council, The Natural Environment Research Council and the British Academy require researchers to offer all research data acquired through research grants to the UKDA and the NERC. The Research Council UK has also announced a revised open access policy in 2012.

- In **Norway**, the Research Council (RCN) has adopted its first policy on open access to research data from publicly funded projects. It provides a general statement on open access policy and, additionally, a more specific statement in the Project Agreement Document regarding archiving requirements, which is the following (Kvalheim and Kvamme, 2013: 20).

"Unless otherwise agreed with the Research Council, copies of all research generated data, including requisite documentation, shall be transferred from the Project Owner to the Norwegian Social Science Data Services (NSD). This shall be carried out as soon as possible and at the latest two years following the conclusion of the project period."

- In **Denmark**, the Danish National Research Foundation states that funded data should be archived at the DDA (Kvalheim and Kvamme, 2013: 20).

- In **Greece**, there are no specific requirements. Greece, as the other EU countries, has to meet the requirements of a limited and flexible pilot action on open access to research data set in Horizon 2020 (European Commission, 2016). All recent research public or EU funding proposals require open data in case of data as final deliverable.

- In **Sweden**, there are no specific requirements. However, applicants receiving grants from the Swedish Research Council must publish their results in open access journals, or archive the article in an open database (Kvalheim and Kvamme, 2013: 20).

- In **Finland**, there are no specific requirements. The Academy of Finland requires that grant applicants provide a Data Management Plan as part of their research plan in accordance with EU' policy lines. It also recommends (and not requires) that (Kvalheim and Kvamme, 2013: 20):

"Academy-funded researchers publish their research articles in open-access electronic scientific journals in cases where there are electronic journals available that meet at least the same quality standards as traditional subscription-based journals"… Academy-funded social science data be delivered to the Finnish Social Science Data Archive (FSD), based at the University of Tampere. Delivery shall take place as soon as possible after Academy funding has ceased."

Moreover, scientific publications often require that the data are archived prior to the publication of the paper.

- In **Netherlands**, organisations carrying out research commissioned by ministries are obligated to deposit data at DANS. This obligation is included in the General Government Terms and Conditions for Public Service Contracts - ARVODI (in Dutch).

- In **Slovenia**, with the purpose of clarifying how research data are treated currently in the country and to propose suggestions for future plans, considering the need for respecting current conditions, MIZS (Ministry of Education, Science and Sports) and ARRS (Slovenian Research Agency) issued a call for the grant on the "Target research programme" with the heading OPEN DATA – Action Plan for the Establishment of a System of Open Access to Publicly Funded Research Data in Slovenia. The project proposal of the Social Science Data Archives (Arhivdružboslovnihpodatkov – ADP) of the University of Ljubljana was accepted for the grant. The Republic of Slovenia has committed, with the recently acquired membership in the OECD international organisation, to follow its Guidelines regarding open access to research data from public funding.

- In **Switzerland**, within the framework of the project Open Research Data Pilot Platform Switzerland (ORD@CH), FORS (lead institution), the Digital Humanities Lab of the University of Basel, and the ETH Scientific IT Services / SIB Swiss Institute of Bioinformatics are developing a publication platform for open research data in Switzerland.

The issue of funding of open data and open access policies is of course of major importance. As far as open access policies are concerned, an interesting perspective arises from the publication of the scientific journal subscription cost in Finland during the years 2010-2015. According to the results of a preliminary analysis (rOpenGov, 2016), Finland paid in total 131.1 million euro subscription and other fees on scientific publishing and the average annual cost was 22 million euro. The data covers all Finnish universities, major public institutions, and some libraries and other institutions. The authors of this document invoked another report, in which it is noted that in 2014,18% of the articles in Finnish universities were published as open access. It was also estimated that publishing all articles as open access would have cost 17 million euro, 5 million euro less than the annual subscription fees paid to publishing houses.

As far as open data are concerned, scientific data infrastructure requires continued budgetary planning and appropriate financial support. OECD (2007) notes that the cost of storing and managing data has been decreased dramatically in recent years. Obviously there is no escaping that maintenance costs money, but paying attention to software sustainability can limit the costs. Maintenance of tools and services is important for a number of reasons: an example is that many results need the original tools to be reproducible. So the need for the establishment of more humanities data centres and for a close knit co-operation between these centres is urgent (Marker, 2015).

In order to conclude this sub-section, we should highlight the following ideas:

1. Considering the domain of historical data, born digital data is becoming increasingly important for historians and humanists.
2. Within CESSDA, some Service Providers collect more historical data than others. A sub-set of SPs that currently archive a considerable amount of historical data are DANS, UKDA and GESIS.
3. When it comes to analyze the data landscape, researchers' attitudes towards data sharing, national's policies and proper infrastructures to manage research data form a major triad.

8. BEST PRACTICES/PRACTICAL ROADMAP

In view of a practical Road Map for CESSDA in order to meet data challenges in our era and given the fact that we should be based and capitalise data potential within CESSDA SP's, most of the best cases presented below are research products or research co-operation of Service Providers. The main reason is that in that way all partners could benefit of these best cases, secondly, we will take advantage as a whole of this recording and thirdly CESSDA strategy and policies for the future should easily take a closer look to these best cases. Needless to say that this type of work can accumulate useful examples and similar activities beyond the end of the project as well since this practical Road Map can serve as a live document for the future.

In particular, the best practice cases concern two of the four data domains to study the official statistics and microdata and big data within health domain. It is about two data domains however the form of data presented is more composite. Health data may also being perceived as big data (volume, velocity, variety) and official statistics and microdata engender administrative and /or operational data as well.

Through our study we quite soon understood and perceived that data domains classifications are more and more interconnected and intersected assisted by increasing progress of technologies, methodologies and knowledge of data collecting, archiving and managing services. More specifically in the case of "Panorama of Census Data 1991-2011" the co-operation agreement between a Research Centre and a NSI opened the way to potential users, researchers, academics and the wider public to easily access and quickly process census data.

In the second-best case example access to micro data has been improved and facilitated through the co-operation of a Service Provider with a NSI in favor of users. In addition to that both parties have accumulated considerable know how for the benefit of potential users and quality of services. The third case concerns the foundation of a new agency in order to collect and disseminate surveys and data bases which are dispersed at the national level through a centralised point. National and international users can easily access ay data production through this hub via specific data catalogues. The forth case provides an example of the future involvement of CESSDA to the newest kind of data, those of big data within health domain, i.e. the collection, storage and curation of MRIs. Within the four cases CESSDA Service Providers act as mediators, knowledge hubs and users' incubators to promote data impact to wider public.

An open question to be addressed in the near future is the need for CESSDA to respond efficiently to the issues arisen of these usages. Issues such as methodologies to comply with significant quantities of data, data protection and ethical issues, users' needs and users training to benefit of these new data domains, co-operation and networking with actors already involved in these domains. European Research Era and increasing data deluge require for social and humanities scientists to actively participate in this landscape under formation. CESSDA by representing more and more European countries thus social transformations and societies all over Europe can efficiently respond to these challenges.

Greece/EKKE

The collaboration prospects between the National Statistical Authorities and the Archives or other research centres are positive, as the exchange of technology and expertise will benefit sides, the broader research community and the public in general. An example of this co-operation in Greece regards the development of the web application **"Panorama of Census Data 1991-2011",** by the The Hellenic Statistical Authority (ELSTAT) and the National Centre for Social Research (EKKE). This long-running co-operation was enshrined in the Memorandum of Understanding signed between the parties in 2012. Thus, the expertise of the Greek National Centre of Social Research and the ability of ESTAT to collect nationally cencuses data resulted in the development of a research tool open to the broader public.

The application was funded by the program "Dynamic management of social databases and cartographic illustrations — SoDaMap" of the Ministry of Education, which is materialised by EKKE in co-operation with the Institute for Information Systems of the Research Centre Athina, in the framework of the National-wide Action "Development Proposals of Research Entities – KRIPIS" of the Ministry of Education. The funding concerns amounts of the National Strategic Reference Framework (NSRF-ESPA) for the period 2007-2013.

The purpose of the application is to enable access, increase analysis potential and permit the easy and detailed mapping of data for the last three censuses (1991-2001-2011). All parties involved aim at satisfying the needs of the research community, but also to increase substantially the number of people who use census data in productive ways. The dynamic nature of access refers to the possibility offered to users to select any combination of variables they want from the relevant Censuses. Furthermore, users are offered the possibility to map census data at various geographical levels, ranging from the Greece's thirteen Regions (NUTS2) to the level of Municipal/Local Communes (more than 6,000 spatial units).

In addition, the application provides access to data and enables mapping within all Greek cities of more than 50,000 inhabitants, and at the level of Urban Spatial Analysis Units (USAU). These units are constituted either by a single Census Section or by more than one adjacent Census Sections, comprising approximately 1,000 inhabitants.

The rationale of this application, as regards the search in the database, is based on: a) selecting, at a first stage, the population the user is interested in (e.g., women aged 40 years and over, having completed tertiary education), and b) extracting tables for the selected population group on the basis of any two variables that the user wants (e.g., profession and branch of employment).

The cartographic part of the application allows for the mapping of variables already incorporated therein. In the near future, the application will also give the possibility to directly map the result of the database search, if of course the search by the user encompasses a geographical reference.

The application will be systematically upgraded with the inclusion of new variables and procedures, thus offering new possibilities. The process of communicating with users is expected to be an important element in the effort to continuously improve the application according to their needs.

Users can access the application at the following link: http://panorama.statistics.gr/

The limitation of access to data of the 2011 Population Census results from the obligation to observe the confidentiality rules governing the provision of statistical data in accordance with Law 3832/2010 of ELSTAT, as in force and European statistical standards. In compliance with the relevant rules, the application does not provide absolute figures, but intervals of values, when the spatial reference level is below the Region (NUTS 2) level.

More detailed data are available on ELSTAT's website, at:
http://www.statistics.gr/portal/page/portal/ESYE/PAGE-cencus2011tables

Slovenia/ADP
Cooperation between ADP and SORS

Partnership between the Slovenian Social Science Data Archives (ADP) and the Statistical Office of the Republic of Slovenia (SORS) intensified intensified with the DwB Project, starting in May 2011[117]. Both organizations recognized the need to concentrate on improving the research environment, access and quality of access to (detailed) official statistics microdata, on the national level.

One of Social Science Data Archives' mission is to provide quality microdata and metadata to different groups of data users: under-graduate students, postgraduate students, young researchers at faculties, research organizations, registered researchers. However, by distributing mostly Public Use Files on ADP catalogue, the needs of experts in the social sciences field, namely registered researchers and research institutes, were not entirely met.

SORS assisted ADP and provided access to the following types of microdata for research: population, housing and agriculture censuses; social surveys; business surveys; and all administrative records SORS uses for the production of official statistics. In return, SORS found the cooperation especially beneficial in the following areas:

- knowledge of DDI metadata standard,
- additional knowledge about students'needs,
- more experience and resources for training of researchers and
- established additional channels for microdata usage promotion.

This intensive partnership between SORS and ADP thus resulted in several fields of cooperation:

- preparation of microdata files for use in SURS's on-site laboratory and remote access facility,
- preparation of non-confidential microdata for less heavy users in ADP's facility,
- introduction of new methods for data protection particularly suitable for students,
- preparation of metadata according to DDI standard with extraction from SURS meta data systems,
- communication of metadata according to DDI standard from both partners,

[117] *Summary and conclusion from official report from the workshop Deliverable D6.3 / Data Without Boundaries project are presented here* (http://www.dwbproject.org/export/sites/default/about/public_deliveraples/dwb_d6-3_regional-workshop-report.pdf).

- cooperation in user training and
- promotion of microdata use: SURS and ADP websites, conferences.

In cooperation with SORS Sector for General Methodology and Standards the anonymisation procedure which follows Eurostat LFS anonymisationcriteria and other advanced methods were introduced in de-individualisation procedures. The analysis of SORS metadata systems and other possible metadata sources was made. ADP counselled and advices SORS work groups on metadata standards. Study descriptions were prepared for a few series of official microdata, ADP DDI extended scheme was used – including methodological, file description, data description, publication, other material etc. metadata fields. Added value was also, that all the required/useful documentation is made available to researchers in one place (codebooks, questionnaires, publications, syntaxes, methodological explanations etc.).

In April 2013 the Statistical Office of the Republic of Slovenia (SORS) hosted the "1st Regional Workshop on Microdata Access in European Countries: Co-operation between National Statistical Institutes & Social Science Data Archives" within the DwB project. The main aim of the workshop was to gather and to connect representatives of National Statistical Institutes (NSI's) and representatives of Social Science Data Archives with the emphasis on the co-operation of these institutions in the Central and Eastern European countries. Participants agreed that the current co-operation makes sense and represents an added value to the development of the national research environment and consequently emphasises a good practice, introducing of which would have a positive influence in other European countries as well.

European Service Centre for Official Statistical Microdata – ESCOS[118]

The concept of a European Service Centre for Official Statistics, ESCOS, was an initiative developed through the DwB project. The ESCOS' central tasks would be to:

- establish a comprehensive web service for microdata from official statistics in Europe, which would include the metadata, routines, discussion forums for research and sharing tools
- promote the use of the official statistics' microdata by hosting training courses, organizing user conferences and by incentivizing research
- improve data access
- support a European Remote Access Network.

Improving data access would involve:

- providing information on access conditions, access sites and application procedures
- supporting expansion of transnational access networks and harmonisation of the conditions of use through Europe
- supporting researchers and the European Statistical System in accreditation.

Thus, the Service Centre could essentially take up the legacy of the DwB and go beyond it. The Integrated European Census Microdata (IECM) database has been vastly expanded as part of the work conducted within the DwB project. The Centre d'Estudis Demogràfics (CED) plans to continue this project and include new samples of census data continuously.

[118] For more information about ESCOS, see the DwB Deliverables, the full report (D5.1 updated v. April 2015).

Sustainability and future development of CIMES after the end of DwB is envisioned in three steps. 1) *Quetelet-PROGEDO diffusion* database continues with its own funds to document additional data sources, including more data sources which can be used for comparative research. 2) Partners from DwB who have participated in Task 5.2 could agree via a Memorandum of Understanding to continue to work on data sources from countries they had worked on prior within DwB. 3) In parallel, it is envisioned that two processes will go on to ensure long term sustainability and development of CIMES. One process will be to discuss with the respective teams the opportunity to develop the tool at the variable level, using MISSY thus going further in the integration of the two tools used within DwB Work Package 5. The process is linked to CESSDA to decide whether and how they can be integrated in the CESSDA work plan.

In the current context, where few CESSDA members have developed co-operation with their respective NSI, and where most NSIs and other government bodies producing official microdata do not provide metadata that could be harvested by the new CESSDA portal to be set up according the CESSDA work plan, CIMES provides a basis for such development in the area of official statistics microdata and can be seen as a 15/17 driver for more co-operation at the national level between the members of CESSDA and the NSIs. The MISSY system is a cornerstone in the services provided by GESIS for official statistics microdata and will be maintained. Based on the current staffing the updating of metadata on EU-SILC, EU-LFS and AES should be secured for the foreseeable future, however currently GESIS has no capacity for continuing the documentation of SES or CIS. While the technical infrastructure which would enable external partners to enter metadata remains in place, none of the DwB partners have the personnel resources to do so without additional funding. However, the technical infrastructure is in place to enable similar co-operation in the future and could be a fruitful possibility for future co-operation in the CESSDA context and a possibility for CESSDA partners besides GESIS to develop expertise on Eurostat Microdata. Such expertise is not only important in providing metadata but also for training and consultancy of researchers and data producers.

Work package 5 of the Data without Boundaries project has produced tools and services to aid the scientific usage of official statistics microdata from Europe. The availability of centrally accessible and structured metadata as well as a wide range of routines will ease the tasks of data exploration and analysis for researchers using European OS microdata. These services benefit not only the research community but are also a considerable aid to data providers. On the one hand the task of structuring and translating metadata is handled for them on the other hand they can also benefit from the increased scientific applicability of their data. As it has been argued for in Deliverable 5.1, we recommend that these services should become an integral part of CESSDA-ERIC ideally by establishing a subunit which is responsible for OS microdata and co-ordinates, maintains and advances these services

The Notion of Circle of Trust

In context of the DwBproject, the notion of a circle of trust[119] was outlined to refer to the need to create circumstances where different parties, such as research data centres and data

[119] Read more and listen to the audio recordings on the EDAF event page under Session 4 - Panel Session: How Useful is the Notion of 'Circle of Trust'? A Vision for the Future.

archives or universities, can rely on each other. Mutual trust is needed for sharing microdata services, i.e. exchanging microdata or providing access to confidential microdata. A concept for basic requirements is necessary for the data providers, and essential for transnational microdata access, but relevant also for the researchers seeking access to data.

When creating a circle of trust, each member joining the circle should be accepted according to the same rules and conditions which are approved by all members. These would cover confidentiality rules and security requirements, but also competence and legal aspects. There would also be set preconditions for the institutions themselves or for technologies providing the access.

Measures needed to gain trust among the actors and to establish a circle of trust:

- shared best practices
- collection and documentation of rules and protocols for transparency
- co-operation agreement
- harmonised contracts for microdata access
- guidelines for the treatment of microdata requests
- catalogue of rules to check which institution can be approved to access microdata
- Researchers' Passport
- security concept and accreditation guidelines for safe centres
- list of security and user demands for a remote access system
- anonymisation concept for scientific use files
- rules and protocols for the transmission of microdata
- guidelines for statistical disclosure methods and output checking, and
- common understanding of responsibilities and similarities.

The notion of Circle of Trust was presented and discussed in the Second European Data Access Forum, EDAF in March 2015. The above description is adopted from the introduction to the concept by M. Brandt (Destatis).

In conclusion, the co-operation with NSIs in general, could gain further momentum with the following obstacles removed:

- NSI's does not see the microdata access for research as part of their mission. The OECD Expert group for international collaboration on microdata access and agreements between Eurostat and CESSDA can foster cultural change in perceived usefulness of research, both for society and as addition of fulfilling the NSI mission (high quality of products, relevance, openness, trust).
- The demand for less elaborate access to microdata needs could bring change. (I.e. few experienced researchers have a privilege of accessing the data, that they gain by devoting a lot of time to get acquainted with the data resources, and who possess the advanced statistical proficiency needed to exhaust full potential of data).
- Thus, training of future researchers and teaching data products could be one of the activities that can gradually rise demand for microdata products.

A report by the "OECD Expert group for international collaboration on microdata access" contains more detailed information on the concept (see Part II, p. 60).

- Change in legislation and practice in implementing it (statistical act, privacy protection), as besides the lack of human resources to devote to microdata access, what NSI refer as a major obstacle to access to microdata is the privacy protection.

Norway/NSD

The Norwegian Model of Cooperation

Norway has been in the forefront internationally developing national research infrastructures, in particular in the social sciences where NSD has been assigned important national infrastructure tasks and responsibilities as early as in the mid-1970s. What makes the Norwegian situation special is the broadly based and formal cooperation between NSD and Statistics Norway, NSD and the National Archive and NSD and the Norwegian Data Inspectorate.

Statistics Norway uses NSD as a mechanism supplying data to the scientific community and the Data Inspectorate uses NSD as a partner implementing the statutes of privacy legislation within the research sector. These are important parts of the foundation on which NSD as national research infrastructure is built. The long-term cooperation between NSD and Statistics Norway and the Data Inspectorate respectively, has undoubtedly contributed to the culture of data sharing in Norway and thus proved instrumental to the conditions under which empirical research operates. It has also demonstrated that it is possible to find solutions protecting the interests in privacy as well as research work. The cooperation with The National Archives since the mid-1980s, which was formalised in an agreement in 2014, has also been important, adding to the legitimacy and trustworthiness of NSD as a national archive for research data.

Trust in the infrastructure is critical for those who fund the infrastructure, those who mandate its use and those who deposit data in it. The close collaborations and partnerships that have developed over four decades have added to the trust in NSD as a national research data archive infrastructure and made it possible for NSD to build a variety of data resources across all fields of the social sciences and increasingly also the health sciences and humanities, such as survey data, regional data, historical data, and social and economic microdata.

The cooperation with Statistics Norway

Statistics Norway (SN) and NSD have both as their obligations to deliver data to social research. There is a close cooperation between NSD and SN and NSD is involved in the delivery of official statistics data to researchers based on a contract with SN.

In Norway, the Statistics Act specifies that data collected for statistical purposes should also be made available for research purposes. Since 1975 NSD has had a formal agreement with Statistics Norway to act as a dissemination channel towards research and education. The general idea is that data from statistical production should be freely available for research through NSD, but the two institutions have to share the workload and expenses involved. The practical consequences are that NSD do the necessary curative work for longer term archiving and prepare data for general dissemination and statistical analysis.

Statistics Norway established a survey division in 1967, primarily to serve its own data collecting purposes, but with capacity also to serve external projects of relevance. This has

proved itself a useful arrangement also for general data collection outside statistics, Statistics Norway presently runs a service marked by flexibility, quality and sophisticated know-how and to a large degree also functions as the preferred data collecting organization for larger research projects. Most important for the data archive, it has supplied the data archive with a steady stream of highly relevant and high-quality survey data. Presently NSD holds about 700 files formally owned by Statistics Norway, but documented, archived and disseminated by NSD. Important to social science research, this is data available free of charge. Of the 700 files, approximately 300 – 350 covers 45 years of Labour Force Surveys, both cross-sectional and panel data. Another 50 covers 40 years of varieties of Surveys of Income and Living Conditions, there are 30+ years of Consumer Expenditure Surveys and Income and Property Surveys. In addition, there is a large variety of other survey material, generally based on a thorough scientific design more than a statistic descriptive need.

This wealth of micro-level sampled data has been coupled with an equally free access to aggregated administrative data. In Norway, the local municipalities are not only the local demographic, political and administrative units, these processes have made them the stable statistical units and NSD has built up a well-documented database covering 200+ years of statistical data production. The database is completed with services for data extraction, harmonization, analysis and visualization.

The general collaborative confidence and understanding of supplementary competencies between the national statistical authorities and the data archive have recently led to a third major collaborative project. As has been pointed out extensively through this report, modern administrative registers make up a potential data source of great relevance for both statistics and research. Over the last 50 years Statistics Norway has developed statistical versions of 5 central registers of particular interest to research, the Population Register, an Education Register, a Tax Register, a Workforce Register and a much-diversified Social Security Register.

In the Scandinavian countries, such data are technically possible to merge because of the personal identification number available since 1963. In Norwegian official statistics, there is a wide use of administrative data (e.g. NAV). Statistics Norway also has close cooperation with administrative registers like the Population Registers and Central Coordinating Register for Legal Entities. The common identifiers (persons and economic units) give SSB the possibility to link information between different sources (administrative and own statistical surveys). Data with identifiers are stored and accumulated over time. This gives theoretically an enormous set of data and this will be a rich data source for research. The data may be combined and presented as cross section data. Data may also be combined in a way that gives longitudinal data.

These sources (administrative data and register data) have not yet been fully integrated in the service package that NSD offers for delivery of official statistics data. In these efforts NSD and SN identified some specific need for development of tools and initiated the RAIRD project and in 2012 the Norwegian Research Council funded a four million € project (RAIRD) aiming at creating a platform for easier and safe remote access to register data. The new access service is a joint venture between NSD and SN. It will provide SN with a more integrated and well documented data storage making it easier to manage their data in the future.

With the new service based on RAIRD technology, Statistics Norway and NSD together establish a national research infrastructure providing easy access to large amounts of rich high-quality statistical data for scientific research. The primary goal is to facilitate and promote high-quality scientific research based on a wide range of high quality statistical data in socio-economic, political and financial areas and by extension, increase the contribution of research to the solution of major knowledge challenges facing society today.

Today Statistics Norway provides access to de-identified data (micro data without identifiers, but with all relevant variables) to researchers from authorised research institutions and specified research projects. The micro data is delivered to the researcher on CD-ROM, and the researcher may analyse the data on his own PC at his own institution. The researchers who are authorised and licensed to have such access to data – will be rather satisfied. There are however some problems with this solution as the general one, caused by the administrative burden of preparing the data and the authorisation of the researcher. This routine sometimes creates a time lag delivering data to the researcher.

The focused collaboration between SN and NSD will increase the total dissemination capacity in this field and make the two institutions by concerted action able to handle an expected increase in requests for data in an efficient way. The aim is to contribute to increased use of quantitative data in health, welfare- and other socio-economic research both nationally and internationally.

RAIRD is a large project that requires technical innovation. Register data are guarded by strong confidentiality requirements, still the ambition is to build an efficient access system that allows analytic use of the data without compromising on data protection. In addition the project needs to develop a user work situation that is regarded as relevant, simple and fast for the user. This has been detailed out as a set of more specific aims:

- User self-service as far as possible
- Absolute protection of data confidentiality
- Integrate data and metadata better than what is now the present practice
- Develop the potential for more efficient and innovative use of the time-dimension of event-based data
- Promote validation possibilities needed in research
- Foster politics of data sharing and open access
- Develop possibilities for interactive analytic work and interactive reporting

The main points developed in the RAIRD project could be summed up as:

1. Development of a data model[120] regarding most data as varieties of event histories. This make it possible to store most types of register data in a unified, but potentially decentralized data store that is easier to maintain, update and version. This data store is directly linked with Statistics Norway's upgraded metadata systems.
2. Development of technologies for interactive use of these data to develop analytic files and analyze the data in a cloud-like virtual research environment behind SN firewalls.

[120] https://statswiki.unece.org/display/gsim/RAIRD+Information+Model+RIM+v1_0

3. A multi-faceted security system primarily based on «the five safes»:
   - Safe users/safe projects; authentication, authorization and access (AAA) at level ¾
   - Safe data/safe environment; Data behind SN firewalls, but no anonymization.
   - Safe output; A specially developed statistical analysis package with automated statistical disclosure control.
1. Work in RAIRD is logged. Work sessions may be reused, edited, rerun, shared, function as documentation of work, etc. But it also indicates that user activity could be administratively documented if need be.
2. The access system is based on a procedure where educational or research institutions register and administer its own users.

RAIRD is a major technical innovation project, building technical solutions of very general usefulness.

France

The definition of Official statistics in France (*cf.* 6.2.1) includes all data productions generated by the statistical surveys of the *National Institute of Statistics and Economic Studies* (**INSEE**). Official statistics also include the data collected by all the organisations with a public service mission. Subsequently, the word "public" is often used instead of "official". Thus, the French official statistical system[121] comprises INSEE and the statistical departments of the ministries involved in statistical operations in their area of expertise. This official statistical system produces quantitative information that constitutes the official data on which public debate is focused.
We stated above that one of the problems at the national scale is the scattered nature of datasets (*cf.* 6.4.1). As many projects need National data, the French Official statistics must be disseminated in a proper way to be used and re-used by the research community.

National Archive of Data from Official Statistics- (ADISP)
The French case is characterised by the multiplication of points of access. In this context, **ADISP**[122] (*Archives de Données Issues de la Statistique Publique*), a data service of the **TGIR PROGEDO**, disseminates surveys and databases produced by INSEE, statistical departments of ministries such as DARES (Directorate for Research, Studies and Statistics of the Ministry in charge of labor, employment, vocational training and social dialogue and economic and social actors), DREES (Directorate of Research, Studies, Evaluation and Statistics, the central administration of health and social ministries), DEPP (Directorate for Evaluation, Foresight and Performance in the fields of education and training) or DEPS (Department of Forecasting and Statistics Studies of the Ministry of Culture and Communication), etc., as well as other public institutions such as CÉREQ (Centre for Research on Education, Training and Employment), or the IRDES (Institute for research and information in health economics). It also disseminates data from international surveys.

Its complimentary free research databases, available for the whole scientific community, were established more than 25 years ago to support some studies conducted by sociologists of the CNRS. The catalogue has been gradually enriched and opened to other disciplines, through

[121] https://www.insee.fr/en/information/2386424
[122] https://www.cmh.ens.fr/greco/adisp_en.php

agreements with new producers among INSEE and other public institutions. ADISP indexes a large collection of more than one thousand surveys, as well as French databases related to social sciences and humanities.

Context

The ADISP's aim is to promote the dissemination of studies and databases produced by the Official Statistics in order to facilitate and enhance the use of statistical data in social sciences. It provides services such as archiving, documentation, formatting of data, assistance for the users, evaluation, and feedback of the data's utilisation by the producers.

ADISP is a department of the large research infrastructure PROGEDO (CNRS). ~~A~~DISP participates in CESSDA and was one of the *Réseau Quetelet* partners (Quetelet network[123]) which is now called *Quetelet Progedo diffusion*.

Quetelet Progedo diffusion, the French Data Archives for social sciences, offers various services and access to different kinds of data to researchers from France and abroad interested in data treatment with the requisite access to databases in the following domains:

- censuses and other databases of French National Statistics;
- major French research data;
- privileged access to international data.

The role of ADISP is significant in linking public statistics and academic system for improving the quality of research.

How ADISP works

The ADISfP team collects and recovers the data from INSEE and other producers. Data will be rework for being validated and improved, in order to be used by the researchers. Within the framework of the agreement between INSEE and ADISP, researchers can get free access to INSEE bespoke tabulations called PSM (Produits Sur Mesure) on more than 20 different INSEE data sources**.** These tabulations are available upon request because the data cannot be accessed through the ADISP catalogue (ADISP is not the owner of those files).

Following a set of dissemination rules, data files available by ADISP depend on their content or on constraints imposed by the different producers. Two kinds of data files are available: "Standard Files" and "Production and Research Files" (FPR). The FPR follow specific rules and must be destroyed after the end of the research activity that initiated their use.

ADISP maintains and expands its catalogue in accordance to agreements signed with INSEE and other public institutions. The catalogue includes different types of data in occasional or regular studies, such as long series (surveys jobs, housing, health), censuses of the population, administrative files, panel data, etc. Upon request, the ADISP team may also provide technical advice on the use of the databases.

Data Access

As we stated before (*cf.* 6.3.2), accessibility is an important dilemma that needs to be alleviated in order to support efficiently academic exploitation of data centres and/or databases. ADISP provides access to researchers (including foreign and Master's students) to use its catalogue while guaranteeing the confidentiality. All files distributed within the framework of Adisp are

[123] http://www.reseau-quetelet.cnrs.fr/spip/?lang=en

freely accessed for research purposes. Any commercial use is prohibited. The use of available files for teaching purposes is prescribed for the majority of files. If a researcher needs to re-use the data, he has to report to ADISP which will inform the producer(s) accordingly. The reporting of the various uses of data to producers aims at facilitating control and promoting the dissemination process.

New Agreements

ADISP proposes the deposit of data of interest to research in social sciences and humanities. Institutions or researchers who possess such data have to respect ADISP' regulations. These regulations are based on agreements guaranteed by the Ministry for Higher Education and Research for institutions, the CNRS and depositing licenses for researchers. ADISP does not own the data that it distributes. It holds and transfers the right to use data in order to promote research in the humanities and social sciences.

International data

ADISP disseminates some French datasets produced in the context of international surveys and databases, such as the microdata edited by Eurostat. At this level, ADISP disseminates the French dataset of the following surveys:

- European Community Household Panel (ECHP);
- European Union Labour Force Survey (EU-LFS);
- European Union Statistics on Income and Living Conditions (EU-SILC);
- Structure of Earnings Survey (SES);
- Adult Education Survey (AES) and Community Innovation Survey (CIS).

ADISP plays an important role in managing the data of the International Social Survey Programme[124] (ISSP). The research unit PACTE (CNRS-University Grenoble Alpes, Sciences Po Grenoble) conducts the fieldwork in France.

Documentation tools

ADISP runs documentation of three websites:

1. Documentation on ADISP Website:

More than 1100 pages of presentation are available. For each survey or database, the website offers a descriptive package including:

2. Documentation of variables on Nesstar-ADISP server

In 2010, ADISP adopted the Nesstar server in order to expand its data catalogue documentation. More than 700 datasets are currently accessible on this server and new data surveys are added continuously.

3. Question bank

The **Question bank** allows users to search into each text, questions, variable names and labels of an important number of surveys disseminated by *Quetelet-progedo-diffusion*. This tool offers the possibility to explore surveys reading the questions asked to the interviewees.

[124] http://www.gesis.org/issp/search-and-data-access/

UK Data Service

Big data / health data Case Study: Curation Challenges for MRI Data

There is growing recognition of the importance of sharing data across the social sciences, and some countries have witnessed great successes in archiving and making available data from surveys, national registers and other data arising from primary research. However, research increasingly extends to include non-traditional data sources and cross-disciplinary studies, so we need to extend our data sharing and curation knowledge.

The past decade has seen a huge rise in studying social phenomena using data not initially collected for research and what we term 'new and novel forms of data', such social media data, online mineable information on the human condition, data from digital sensors, financial transactions and administrative records. At the UK Data Service, we have started to acquire these 'new' types of data that fall outside of our traditional disciplinary focus. Equally we have started to receive new types of data that social sciences researchers generate during their research.

Neuroimaging data is just one example, typically resulting from monitoring brain activity in psychological, behavioural and linguistic research. More and more researchers are making use of Magnetic Resonance Imaging (MRI) in their primary research. Psychologists in Japan are using MRI scans to map where happiness emerges in the brain[125]. Similarly, social scientists and biologists from California are using MRI scans to explore questions about how basic social rewards are processed in the brain, and psychologists in Cambridge have used MRI to show antisocial behaviour in the brain[126] [127]. Thus, as psychological or psychiatric studies interface more with traditional social science disciplines, archives will see an increase in the amount of 'non-traditional' data of this type being offered for curation.

Clinicians also argue that sharing neuroimaging data may lead to a better understanding of the brain, and hopefully advances in clinical diagnosis and treatment of neurological and psychiatric disease. As the process is so expensive, duplication of effort could be avoided, replication and validation could be conducted and data quality could be collectively assessed.

MRI data can be large, complex and have usability challenges, plus they will typically be accompanied by experimental, observational or behavioural data of the people studied. Investigating how to assess, document and curate them for future research is important. Based on existing domain knowledge in the neurosciences and from MRI datasets we have recently received[128], this case study presents recommendations for long-term data curation and access.

Key challenges for curation

Information generated from neuroimaging results in complex data that can be arranged and analysed in many different ways. Until recently there has been no standard ways to organise, share or preserve neuroimaging data and there are a multitude of data formats, pre- and post-

[125] https://www.nature.com/articles/srep16891#abstract
[126] http://conte.caltech.edu/
[127] http://onlinelibrary.wiley.com/doi/10.1111/jcpp.12581/abstract
[128] Just 1 example dataset: Pernet, Cyril, Gorgolewski, Krzysztof and Ian, Whittle (2017). A neuroimaging dataset of brain tumour patients. [Data Collection]. Colchester, Essex: UK Data Archive. 10.5255/UKDA-SN-851861

processing techniques, analytical processes, tools and platforms. More recently solutions have been offered by initiatives such as **OpenfMRI** and the **International Neuroinformatics Coordinating Facility (INCF)** which can help archivists prepare raw and pre-processed MRI data for long term availability and reuse. There have also been some excellent papers written by neuroscientists, for example as part of the organisationfor Human Brain Mapping (OHBM) around best practices in analysing and sharing MRI data[129, 130].

MRI scanning and data collection

MRI is a medical imaging technique to create pictures of the anatomy and physiological processes of the human body. An example is a brain scan which can be used to detect various disorders, such as tumours or aneurysms. It does not use x-rays as in other radiology techniques, and is thus non-invasive. In addition to use for medical purposes, it is also utilized in research to measure brain structure and function which makes use of two different types of scanning: *structural MRI* which views the anatomical structure; and *functional MRI (fMRI)*which views the metabolic function by measuring blood or oxygen flow. To help think about the images, MRI has a high spatial resolution and fMRI has a much longer temporal resolution.

MRI scans can be:

- 2-D, a single cross-section of the brain
- 3-D, multiple cross-section scans aggregated into a 3D model or 'volume' of the brain)
- 4-D, such as many 3D volumes captured over time.

In 3D scans tissue is mapped and represented as 'volume elements' or voxels (like a pixel for conventional digital images) which each representing a value in three-dimensional space. In the case of fMRI the measurement of changes in these values invoked by a stimuli or psychological task facilitates a range of statistical analysis techniques.

MRI Formats

MRI data formats come in two types: formats that standardise the images generated according to medical diagnosis, e.g. **DICOM**; and formats that are created to facilitate post-processing analysis, e.g. **Analyze**, **NIfTI**, **MINC**.

MRI data files are typically stored using either: a single file that contains metadata and image data, with metadata stored at the start of a file (e.g. **DICOM**, **Minc** and **NIfTI**); or a single file that stores metadata and a separate file that contains the image data (e.g. Analyze which uses two-files (.hdr and.img).

As an example, a **DICOM** file consists of a header and image data sets within a single file[131]. The header information comprises standardised information on patient demographics and study parameters. Once the data has been transferred onto a local machine for processing the

[129] Nichols,T.E, et al. (2016) 'Best Practices in Data Analysis and Sharing in Neuroimaging using MRI' http://www.humanbrainmapping.org/files/2016/COBIDASreport.pdf
[130] Poline, J. et al (2012) 'Data sharing in neuroimaging research' https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3319918/
[131] Varma, D. (2012) 'Managing DICOM images: Tips and tricks for the radiologist' https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3354356/

DICOM header is removed. **DICOM** format files are taken from the scanner then converted to the **NIfTI** format, a standard defined by the Neuroimaging Informatics Technology Initiative to act as an image data interchange format between neuroimaging programs and analysis packages. All major image processing software packages can convert from **DICOM** to **NIfTI**.

NIfTI files (preferably compressed due to large image file sizes) are used by the majority of public domain processing packages and is the pre-processing format recommended by the Brain Imaging Data Standard (BIDS).[132]

Analyze format data has been superseded by **NIfTI**and there are common tools that can easily convert **Analyze** and **MINC** (Medical Imaging file format) files to **NIfTI**.

MRI Storage

Typical scanners do not usually have large local storage capability nor do they offer a secure setting and, as such, scanners are generally integrated with a Picture Archiving and Communication System (PACS) which provides long-term image storage and retrieval. Pre-processed MRI data are stored in a PACS in the **DICOM** format, a standardised medical image format that facilitates interoperation between scanning technologies and platforms.

In general, analysis packages do not read **DICOM** format files thus in order for MRI data to be processed it has to be converted to a standard format for analysis.

Which data?

One key data sharing challenge from any research is deciding *which* data are to be shared, taking into consideration ethical and data protection issues. Various versions of data may offer different analytic potential:

- **DICOM** raw files from the scanner - offers greater analytic power to users
- **NIfTI** pre-processed converted data – facilitates instant analysis
- fMRI data processed data – provides more information than can be conveyed in a static image
- processed statistical summary maps for the whole study- offer more holistic information

Personal, behavioural or attitudinal attributes collected about studied individuals should be shared along with the data, bearing in mind the ethical considerations discussed below.

Ethical Considerations

As in any human subject research, the rights of the 'scanned' subjects need to be respected and consistent with the study's ethical review process; and bear in mind the different rules and ethical regulations between countries. The Open Brain Consent project provides some useful sample consent forms and information sheets written with data sharing in mind[133].

[132] Brain Imaging Data Standard (BIDS)http://bids.neuroimaging.io/
[133]https://open-brain-consent.readthedocs.io

Prior to making MRI data available, some degree of de-identification will need to be carried out to protect the privacy of the subject. For the images themselves, because features like the eyes, nose, and mouth become relevant when recognising a familiar individual, techniques in the literature propose a technique known as 'defacing', which essentially involves using algorithms to disguise full facial features.[134]

Any identifying attributes or medical details can be completely removed or generalised from image headers, **DICOM** files and file pathnames, a technique known in this domain as 'scrubbing'.[135] An example is providing age instead of actual birth date.

Finally, other data collected for research that is related to the subject needs to be considered. The archivist must decide for these which level of access is required to meet relevant ethical and legal considerations. Removing the linkage key is one way or housing disclosive data under more restrictive conditions another.

Organising data

Neuroimaging experiments result in complex data that can be arranged and analysed in many different ways. Of interest are the solutions offered by initiatives such as OpenfMRI and the International Neuroinformatics Coordinating Facility(INCF) which offer advice on how we can prepare raw and pre-processed MRI data for long term availability and reuse.

Organising files and directories and using naming conventions are useful as these facilitate automation of processing and analyses. The Brain Imaging Data Structure (BIDS)offers a standardised data structure for organizing and describing MRI data providing a detailed directory hierarchy for images, plus the use of plain text files for noting key information about the dataset, such as its provenance. It proposes using file formats and metadata that are compatible with common data analysis software (such as OpenfMRI.org, LORIS, COINS, XNAT and SciTran), and offers online validation of dataset integrity. Analysis workflows and scripts should be documented (e.g. as a shell script, Matlab or Python) and a Readme.txt file which explains how to execute the workflows/scripts is also recommended. By making explicit the organisation of data files and workflows research replicability is better enabled.

Documentation

To make MRI data findable and reusable for secondary analysis it needs to be interpreted within the context of its origins, in its generation and analyses; thus sufficient description is recommended. Archival staff should ensure that as much detail as possible is provided in the documentation accompanying the data for deposit, the standard DDI type metadata, together with documentation about the data collection and processing actions on the data, plus a read file (preferable machine readable) with instructions on the formats and what software are needed to read them.

[134] Bischoff-Grethe et al (2007) A Technique for the Deidentification of Structural Brain MR Images https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2408762/
[135] Nichols, T.E, et al. (2016) 'Best Practices in Data Analysis and Sharing in Neuroimaging using MRI' http://www.humanbrainmapping.org/files/2016/COBIDASreport.pdf

The Neuroimaging Data Model (NIDM) further defines standardised metadata elements for the domain of human brain mapping, using provenance information to link information about different stages of the research process from dataset descriptors and computational workflow, to derived data and publication[136].

Useful data documentation for neuroimaging would include:

- Demographic details and how subjects were identified, selected and consented
- Experiments, tasks or stimuli applied during the scanning
- Length of scanning session and duration of each run
- Instrumentation and software, e.g. make, model, field strength of the scanner etc.
- Details of each manipulation including name of software, tools and operating systems involved in the analysis
- 'Provenance trace' of the analysis including the software and pipelines (some analysis involves multiple tools and platforms) describing the workflow connecting these tools to aid reproducibility and replicability
- Processing methodology detailing the process of extracting results from data, distilling down vast datasets to meaningful and understandable statistical summaries (model fitting followed by statistical inference or prediction)

Storage and processing at scale: the future for curating big data

New technologies allow for the integration of a diverse range of differing and sometimes competing ontologies and metadata schema that are produced from MRI data. Storing data in new database models like RDF mean that we can accommodate data of vastly different types in a single flattened schema; think of this as a web of data points akin to the web of documents that was revolutionary around thirty years ago. This hyperlinked and interconnected model for data assists with tasks like harmonisation across different file formats and discovery of data and outputs by enabling instant semantic linkage between any data point, variable or case. It's like making a new chemical compound by precision bonding the atoms in two molecules, rather than mixing the contents of two test tubes and hoping that the ratios are correct. An example is the ability to leverage the link between data that captures the characterisation of pulmonary COPD from MRI scans, self reported COPD, or from medical records and markers in genomic analysis from metadata records.[137] Traditionally, linkage between these disparate sources would be constrained by their physical formats. Applying a common "flattened" universal format before analysis means that we analyse all of this data as a single logical entity.

Summary recommendations

- Encourage researchers to use the BIDS standard for organizing and describing MRI data, procedures and tools. This can ensure interoperation and facilitate data sharing and re-use.
- Store raw pre-processed MRI data direct from the scanner in the standard **DICOM** format, a standardised medical image format that facilitates interoperation between scanning technologies and platforms.

[136] Neuroimaging Data Model (NIDM) http://nidm.nidash.org/

[137] Hoffman, E.A et al (2015) Pulmonary CT and MRI phenotypes that help explain chronic pulmonary obstruction disease pathophysiology and outcomes http://onlinelibrary.wiley.com/doi/10.1002/jmri.25010/full

- Store pre-processed data in the standard **NIfTI** format (preferably compressed due to large image file size), a format used by most processing packages.
- Ensure that data sharing is in alignment with suitably crafted ethics and consent documents which accompany the data.
- Ensure that the data has been appropriately anonymised and de-faced and that all sensitive personal information is has been withheld or obscured.
- Ensure that adequate contextual documentation accompany the MRI data in order for it to be understood and re-useable.
- Encourage researchers to adopt the Neuroimaging Data Model (NIDM) to describe the 'provenance trace' of the analyses.
- Consider using technologies that allow for the integration of a diverse range of ontologies, metadata schema and data formats.

9. CONCLUSION

Today CESSDA has just become CESSDA ERIC, opening up new possibilities of co-operation and networking, whilst widening and expanding CESSDA membership is in the core of future strategy. CESSDA has become a mature network, able to capitalise knowledge, expertise needed to deal with archiving and managing of data produced/provided at national and European level. At the same time, the data landscape has been widened in terms of volume and plurality due to the emergence of technological advances, new types of data and new actors appeared at the European and international scene. Main research findings are based on CESSDA SPs' activities, trends and Open Access policies worldwide and current as well as new types of data located in and outside CESSDA SP's. The data landscape as depicted within the frame of this particular task may contribute to the on-going CESSDA ERIC strategy and policies as well. Despite the fact that we studied four particular data domains, namely academic, health, historical and official statistics/big data, we came up to certain concluding remarks that could serve to reflect upon CESSDA ERIC in the near future.

In particular, for the academia domain despite the vast number of datasets provided by CESSDA SPs, the emergence of new types of data and data producers, has led to the understanding that CESSDA archives need to diversify their content soon. Survey data collected by academics and researchers have driven the activities of social science data archives for a long time. Nevertheless, types of data such as big data and microdata may place different demands on the archive's technical infrastructure, compared to sample surveys and, more generally, quantitative statistical data. We should also stress out the movement from traditional diplomatic, economic, and political history toward newer approaches embracing social and oral history, as well as the increasing importance of time series within the discipline. Issues of classification and terminology regarding historical data should be of importance for the SPs, in order to offer easy access to the research community. Official statistics data remains a considerable pool for the research /academic community. It is a moving landscape filled with obstacles already present on a national level. CESSDA can capitalise past experience (DwB project etc.), long expertise on data protection and procedures, as well as best practices at country level (France, Nordic countries cooperation) in order to meet increasing researchers' needs on different types of OS data.

Amongst main findings of this report, some issues have arisen to be further discussed within CESSDA concerning its strategic planning for widening data perimeter. Two main broad areas of interest came up, namely the increasing and widening of data collection provided on the one hand, and the improvement of the existing or the development of new tools regarding data curation, elaboration and dissemination on the other hand. These two broad areas of interest are not contradictory, as many SPs, especially the biggest ones, are involved in both areas. It is understandable that those strategies are not equally feasible for the smaller SPs, which have to take into account financial and human resources limitations. At the same time, researchers' needs are increasing across Europe (and the globe). Thus, the need of achieving economies of scale through networking, cooperation and exchange of expertise amongst the SPs, as well as with other actors outside CESSDA, seems meaningful in order to ensure further development and sustainability. The provision and curation of qualitative data for example (including historical, academic or health data) is a challenging task regarding technical, legal and ethical

aspects. The development of relative strategies and tools could be taken over by some SPs that could build upon and further share knowledge and expertise within CESSDA and for the benefit of the academic and research community and broader public.

Another case is the emergence of the importance of big data, which constitutes one of the major challenges for SSH. Dealing with Big Data in social sciences mean also dealing, at a great extent with people's perceptions, life histories etc. Given that for the years to come the combination of social sciences analysis and techniques and computational sciences to further develop and promote data content for policy driven implications must have in view inclusive and participatory societies, CESSDA and SPs can be actively involved. Within big data, the domain of health seems also quite promising. A limited for the time being number of SPs has started dealing with big data within health domain. Moreover, SPs can act as mediators or significant interlocutor between public agencies like NSIs, ministries or university departments in order to provide services, build networks and being a reliable partner, as it is the case of ADISP. In the appendix 1, a list of actors operating in different data fields has been integrated. The appendix can be used as a living document mapping potential partners, locating potential "competitors" whilst the "panorama" of their activities can influence CESSDA activities and strategies in the future.

With regard to the specific data domains that have been explored in this report, future strategies should promote the establishment of long term collaborations with actors e.g. other RIs or organisations within European era that collect or provide data that are of interest to CESSDA. These agreements or collaborations should take into account both constant development of technological advances as well as research progress in order for CESSDA to meet users 'needs. This is necessary, as the emergence of new data types and the increasing volume of available data.,

CESSDA's maintenance at the forefront presupposes that SPs should make bridges with actors at the national level seeking networking for ensuring data flow from various sources. Bridges that fill gaps and strengthen CESSDA position further. In particular, the assets of CESSDA' SPs regarding the content of the data collections and data management can even lead to a new sharing of labour among main stake holders in the field. In other words, the openness to the wider environment can be an important component of CESSDA strategic plan in the years to come. CESSDA and SPs operate within the broader European ecosystem and they have to maintain links and build bridges with important key actors. Cooperation would also allow collective and more effective work in many issues arising during the exploration of the specific data domains of this report, such as legal and ethical aspects regarding the handling of sensitive data, the development of more coherent metadata descriptions or the elaboration of classification and documentation standards.

REFERENCE LIST

Beck, U. (1992). *Risk society*. London: Sage.

Borg, S. (2014). Open access to research data. *FSD Bulletin*, 40, 02/14. http://www.fsd.uta.fi/lehti/en/40/

Chan, L. et al (2002) Budapest open access initiative. Retrieved from http://www.budapestopenaccessinitiative.org/

Chen, H., Chiang, R.H.L. & Storey,V.C.(2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, Vol. 36, (4), pp. 1165 – 1188.

Cornuau, F. & Silberman, R. (2015). *Researchers' needs: understand how they work to implement a EU-RAN*. European Data Access Forum, Luxembourg, 24th – 25th of March 2015. Retrieved from http://www.dwbproject.org/export/sites/default/events/doc/edaf2_files/dwb_edaf2_s2panel-intro_researchers-needs-2_cornuau-silberman.pdf

Council of the European Union (2016). *The transition towards an open science* system. Council *conclusions*. Retrieved from http://data.consilium.europa.eu/doc/document/ST-9526-2016-INIT/en/pdf

Council of the European Union (2015). *Draft Council conclusions on open, data-intensive and networked research as a driver for faster and wider innovation*. Retrieved from http://data.consilium.europa.eu/doc/document/ST-8970-2015-INIT/en/pdf

Digital Curation Centre - DCC (2016). *Overview of funders' data policies. Web presentation*. Retrieved from http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

Directors General of the National Statistical Institutes – DGINS (2013) *Scheveningen Memorandum. Big Data and Official Statistics*. Retrieved from https://ec.europa.eu/eurostat/cros/system/files/SCHEVENINGEN_MEMORANDUM%20Final%20version.pdf

Economic and Social Research Council - ESRC (2015). *ESRC research data policy*. Retrieved from http://www.esrc.ac.uk/files/about-us/policies-and-standards/esrc-research-data-policy/

European Commission (2016). *Guidelines on data management in Horizon 2020*. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

European Commission (2016a). *AGA – Annotated model grant agreement. H2020 Programme*. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf#page=215

European Commission (2016b). *Guidelines on open access to scientific publications and research data in Horizon 2020*. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

European Commission (2016c). *Access to and preservation of scientific information in Europe*. Report on the implementation of Commission Recommendation COM (2012), 4890 final. Retrieved from http://ec.europa.eu/research/openscience/pdf/openaccess/npr_report.pdf#view=fit&pagemode=none

European Commission (2015). *Validation of the results of the public consultation on Science 2.0: Science in* Transition. https://ec.europa.eu/research/consultations/science2.0/science_2_0_final_report.pdf

European Commission (2013). *Final report on the public consultation on open research data*. Retrieved from http://ec.europa.eu/digital-single-market/node/67533

European Commission (2013a). *Implementing Regulation (EC) No 223/2009 of the European Parliament and of the Council on European Statistics as regards access to confidential data for scientific purposes and repealing Commission Regulation (EC) No 831/2002*. Retrieved from http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32013R0557

European Commission (2012a). *Commission recommendation on access to and preservation of scientific information*, COM (2012) 4890 final of 17.7.2012. Retrieved from https://ec.europa.eu/research/sciencesociety/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf
European Commission (2012b). *Towards better access to scientific information: Boosting the benefits of public investments in research*. COM (2012) 401 final of 17.7.2012. Retrieved from https://ec.europa.eu/research/science-society/document_library/pdf_06/era-communication-towards-better-access-to-scientific-information_en.pdf
European Commission (2011). *Open data. An engine for innovation, growth and transparent governance*. COM(2011) 882 Final. Retrieved from http://www.europarl.europa.eu/RegData/docs_autres_institutions/commission_europeenne/com/2011/0882/COM_COM(2011)0882_EN.pdf
European Commission (2010). *Digital Agenda for Europe*. COM (2010) 245 final. Retrieved from http://eur-lex.europa.eu/legal-content/BG/ALL/?uri=uriserv:si0016
European Commission (2007). *Scientific information in the digital age: access, dissemination and preservation*. COM (2007) 56 final of 14.2.2007. Retrieved from http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52007DC0056
European Statistical System (2014). *ESS Vision 2020 Building the future of European Statistics*. Retrieved from http://ec.europa.eu/eurostat/documents/42577/6906243/ESS+vision+2020_V2/35911206-3968-4548-adcc-882c797d9ca4
Eurostat (2016). *Access to microdata for scientific and statistical purposes*. Retrieved from https://www.ecb.europa.eu/pub/conferences/shared/pdf/160705_8th_stats_conference/Kotzeva.pdf?40e9e00bd758c23663be6634442d2b36
Expert Advisory Group on Data Access (2015). *Governance of data access*. Retrieved from https://wellcome.ac.uk/sites/default/files/governance-of-data-access-eagda-jun15.pdf
Fält, K. (2015). Open up your data, humanist! *FSD Bulletin*, 43, 03/15. Retrieved from http://www.fsd.uta.fi/lehti/en/43/
Franzmann, G. (2015). The online database Histat as an example for research-promoting infrastructures for studies in quantitative historical research. *Economies et societes, Serie"Histoire economique quantitative"*, 50, pp. 821-856.
Kondyli, D., Tzortzopoulou, M., Vezyrgianni, K. (with contributions from A. Cornilleau, M. Cros, R. Silberman and P. Tubaro in Section 3, and C. Kappi in Part II. (2012). *Data collection strategies: CESSDA organisations and their relation to data collections outside CESSDA (D10.5a)*. Retrieved from http://ppp.cessda.net/doc/D10.5a_Audit_collection_strategies.pdf
Kvalheim, V., Kvamme, T. (2013). *Policies for sharing research data in social sciences and humanities*. IFDO. Retrieved from http://ifdo.org/wordpress/wp-content/uploads/2015/07/ifdo_survey_report.pdf
Järvelä, K. (2015). FSD probes researcher attitudes towards open access to data. *FSD Bulletin*, 43, 03/15. Retrieved from http://www.fsd.uta.fi/lehti/en/43/researcher_attitudes.html
Järvelä, K. (2015a). Humanists worry over archiving ethics, data ownership and practical issues. *FSD Bulletin*, 43, 03/15. Retrieved from http://www.fsd.uta.fi/lehti/en/43/humanists_worries.html
Johnsen, Å. (2005). What does 25 Years of experience tell us about the state of performance: Measurement in public policy and management? *Public Money & Management*, 25, pp. 9-17.
Mandinach, E.B. (2012). A perfect time for data use: Using data-driven decision making to inform practice. *Educational Psychologist*, Vol. 47, (2), pp.71-85.
Marker, HJ. (2015). The importance of data infrastructures for the humanities. *FSD Bulletin*, 43, 03/15. Retrieved from http://www.fsd.uta.fi/lehti/en/43/column_data_infrastructures.html
Max Plank Society and Max Planck Institute for the History of Science (2003). *Berlin declaration on open access to knowledge in the sciences and humanities*. Retrieved from https://openaccess.mpg.de/Berlin-Declaration

Mergel, I., Rethemeyer, RK. & Isett. K. (2016). Big data in public affairs. *Public Administration Review*, 76 (6), pp.928-937.
National Documentation Centre (2016). *Open access*. Retrieved from http://openaccess.gr/openaccess/evolution.dot
Niemeijer, D. (2002). Developing indicators for environmental policy: data-driven and theory-driven approaches examined by example. *Environmental Science & Policy*, 5 (2), pp. 91–103.
OECD (2014). *OECD's Expert Group for international collaboration on micro-data access*. Retrieved from https://www.oecd.org/std/microdata-access-executive-summary-OECD-2014.pdf
OECD (2013). *New data for understanding the human condition. International perspectives.* Retrieved from http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf
OECD (2007). *OECD principles and guidelines for access to to research data from public funding*. Retrieved from http://www.oecd.org/sti/sci-tech/38500813.pdf
OECD (2005). *Glossary of statistical terms*. Retrieved from https://stats.oecd.org/glossary/detail.asp?ID=1656
OECD (2004). *Declaration on access to research data from public funding*. OECD C(2004)31/REV1, 30 January 2004. Retrieved from: http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=157
Priddy, M. & Wittenberg, M. (2015). *What researchers want… from a resource discovery service for OS microdata.* European Data Access Forum, Luxembourg, 24th – 25th March. Retrieved from: http://www.dwbproject.org/export/sites/default/events/doc/edaf2_files/dwb_edaf2_s2panel-intro_researchers-needs-1_priddy.pdf
Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, 41(5), pp. 885–899. https://doi.org/10.1177/0038038507080443
SERISS (2017). *WP6: New forms of data – legal, ethical and quality issues*. Retrieved from https://seriss.eu/about-seriss/work-packages/wp6-new-forms-of-data-legal-ethical-and-quality-issues/
Sotiropoulos, D. (2006). *Survey for the higher educational system in Greece*. Athens: ELIAMEP. Retrieved from www.eliamep.gr
Struijs, P. Braaksma, B. & Daas, P. JH (2014) Official statistics and Big Data. *Big Data & Society*, April–June, pp. 1–6.
Suber, P. (2015). *Open access overview*. Retrieved from: http://legacy.earlham.edu/~peters/fos/overview.htm
Suber, P. et al. (2003). *Bethesda Statement on open access publishing*. Retrieved from http://legacy.earlham.edu/~peters/fos/bethesda.htm
rOpenGov (2016). *Scientific journal subscription costs in Finland 2010-2015: a preliminary analysis.* Retrieved from http://ropengov.github.io/r/2016/06/10/FOI/
Tanaka, S. (2015). Preconceiving pasts in a digital age. *HISTOREIN*, Vo. 15 (2), pp. 22-29.
Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M. & Frame, M. (2011). Data sharing by scientists: Practices and perceptions. *PLoS ONE* 6 (6): e21101. doi:10.1371/journal.pone.0021101
Tubaro, P., Cros, M. & Silberman, R. (2012). *Transnational access to official microdata and accreditation in Europe. State of the art and challenges ahead*. 1st European Data Access Forum, 27 March 2012. Luxemburg.
UN Statistics Office (2014) *Fundamental Principles of Official Statistics* (A/RES/68/261 from 29 January 2014). Retrieved from http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx
Van Den Eynden, V. & Bishop, L. (2014). *Incentives and motivations for sharing research data: researcher's perspectives*. Essex: UK Data Archive, University of Essex. Retrieved from http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf

List of Tables:

List of Pictures:

APPENDIX I

A. HEALTH DATA – STATE OF PLAY/OTHER ACTORS

UNITED KINGDOM.

National Statistics Record Matching. http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/gss-methodology-series--25--methods-for-automatic-record-matching-and-linkage-and-their-use-in-national-statistics.pdf
Health Administrative Data: Exploring the potential for academic research [2010]. http://www.adls.ac.uk/wp-content/files_flutter/1295883198ADLSHealthResearchpaper.swf
Office for National Statistics Longitudinal Study. https://www.ucl.ac.uk/celsius/about-the-ls/what-data-does-the-ls-contain
Scottish Longitudinal Study. http://sls.lscs.ac.uk/about/ and http://sls.lscs.ac.uk/guides-resources/what-data-are-included/health-events/
Scottish Health Data. http://www.adls.ac.uk/wp-content/uploads/Introduction-to-ISD-administrative-data.pdf
Northern Ireland Longitudinal Study. http://www.qub.ac.uk/research-centres/NILSResearchSupportUnit/About/WhatistheNILS/ and http://www.qub.ac.uk/research-centres/NILSResearchSupportUnit/About/
A Health and Biomedical Informatics Research Strategy for Scotland. http://www.gov.scot/Resource/0047/00475145.pdf
The Information Governance Review. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/192572/2900774_InfoGovernance_accv2.pdf
Strengthening the UK's capability in health informatics research. http://masoninstitute.org/wp-content/uploads/2013/08/Conference-Report-HIRC-launch-May-2013.pdf
Administrative Health Data. http://www.adrn.ac.uk/catalogue#facet6
UK Data Service. https://www.ukdataservice.ac.uk/get-data/themes/health
Enhancing discoverability of public health and epidemiology research data [2014] – Summary report.
http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp056925.pdf and Full Report http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp056916.pdf with annexes http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtp056915.pdfhttp://www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Public-health-and-epidemiology/WTP056917.htm
Access – public attitudes - http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_grants/documents/web_document/wtp060244.pdf and http://www.wellcome.ac.uk/stellent/groups/corporatesite/@msh_grants/documents/web_document/wtp060243.pdf

https://www.mrc.ac.uk/documents/pdf/the-use-of-personal-health-information-in-medical-research-june-2007/
Other:
http://www.esrc.ac.uk/files/research/administrative-data-taskforce-adt/better-information-means-better-care-geraint-lewis/

SWEDEN.

LifeGene. https://www.lifegene.se/For-scientists/
Register data (in Swedish). http://www.registerforskning.se/en_us/
SIMSAM- Swedish Initiative for Research on Microdata in the Social And Medical Sciences. http://simsam.nu/publications/
Swedish Registers- A Unique Resource for Health and Welfare (full text). http://simsam.nu/wp-content/uploads/2013/04/simsam_booklet_eng.pdf
The National Board of Health and Welfare. http://www.socialstyrelsen.se/statistics
Swedish Cohort Consortium: http://cohorts.se/
The Swedish personal identitynumber: possibilities and pitfalls in healthcare and medical research.
http://www.socialstyrelsen.se/publikationer2011/theswedishpersonalidentitynumber-possibilitiesandpitfallsinhealthcareandmedicalresearch
Swedish national inpatient register. http://www.socialstyrelsen.se/publikationer2011/externalreviewandvalidationoftheswedishnationalinpatientregister

AUSTRIA

Statistics Austria. Holds data on aggregated and individual level health data.
http://www.statistik.at/web_en/statistics/index.html

DENMARK

Healthcare Denmark. http://healthcaredenmark.dk/
National Strategy for Data Management. https://www.deic.dk/sites/default/files/uploads/PDF/National%20Strategi%20for%20Forskningsdata%20Management%202015-2018.pdf
Statistics Denmark. http://www.danmarksstatistik.dk/en/TilSalg/Forskningsservice
Statens Serum Institut. Research on infectious diseases and biological threats as well as control of congenital disorders http://www.ssi.dk/English.aspx

FINLAND

Kelasto Database. Data on drug use, sickness leave and rehabilitation.
http://www.kela.fi/web/en
Sotkanet. National Institute for Health and Welfare information service offers key population welfare and health data from 1990 onwards on all Finnish municipalities.

https://www.sotkanet.fi/sotkanet/en/tietoa-palvelusta
Statistics Finland. http://www.stat.fi/til/ter_en.html

GREECE

Health Atlas. https://healthatlas.gov.gr/Statistics/#!/
Hellenic Pasteur Institute. http://www.pasteur.gr/?lang=en
Hellenic Food Authority. http://www.efet.gr/portal/page/portal/efetnew/library/plans
Hellenic Statistical Authority. http://www.statistics.gr/en/home/
Institute of Child Health. http://www.ich.gr/en/
Hellenic Centre for Disease Control & Prevention. http://www.keelpno.gr/en-us/home.aspx
National School of Public Health. http://www.nsph.gr/?page=home&lang=EN
University Mental Health Research Institute. http://www.ektepn.gr/en/index.htm
National Agency for Transplants. http://www.eom.gr/
National organisationfor Medicines. http://www.eof.gr/web/guest
Athens Medical Society. http://www.mednet.gr/
The National Centre for Social Solidarity http://www.ekka.org.gr/EKKA!show.action?lang=en

GERMANY

Federal Bureau of Statistic. Demographic data on birth and fertility rates, mortality, life expectancy and population-related data on the causes of death.
https://www.destatis.de/EN/Homepage.html;jsessionid=55E3FC0F1F0AC4C7B8D98A6DF9FD750F.cae2
GEKID. Association of Population-based Cancer Registries in Germany.
http://www.gekid.de/index_e.html
Robert Koch Institute. Epidemiological centre for infectious diseases.
http://www.rki.de/DE/Home/homepage_node.html

NETHERLANDS

Elexir-NL. Data infrastructure for the life sciences.
https://www.dtls.nl/elixir-nl/elixir-nl-2/
Health. Research infrastructure inpersonalised medicine and health research.
https://www.health-ri.org/

NORWAY

Norwegian Institute of Public Health. https://www.fhi.no/en/
Statistics Norway: Registers and statistics of Causes of death, disabilities, health conditions and living habits and health services. https://www.ssb.no/en/helse
Norwegian Prescription Database.
https://www.fhi.no/en/hn/health-registries/norpd/norwegian-prescription-database/

FRANCE

Institut des données de santé
http://www.institut-des-donnees-de-sante.fr/
MSSH-EHESP – Maison des sciences sociales du handicap

http://www.bdsp.ehesp.fr/reseau/mssh-ehesp/
Échanges de données dans l'espace sanitaire et social
http://www.edess.org/joomla/index.php
INSERM - Banque d'information sur les recherches de l'INSERM (http://bir.inserm.fr/)
http://www.inserm.fr/
Delfodoc
http://aphp.aphp.fr/ressourcesdocumentaires/base-documentaire-delfodoc/
Institut Open Health
www.openhealth-institute.org
RESSAC : RÉseau Santé Social d'Administration Centrale
http://ressac.sante.gouv.fr/exl-php/cadcgp.php?CMD=CHERCHE&MODELE=vues/masts_internet_consult_page_accueil/tpl-q.html&query=1&TABLE=ILS_DOC&NOMFONDS=Cadic%20Int%E9grale&NONVALID=
Prisme http://www.documentation-sociale.org/
Centre de Documentation de l'AP-HP http://aphp.aphp.fr/ressourcesdocumentaires/
Centre national de documentation audiovisuelle en santé mentale (CNASM) http://cnasm.fr/
Eco-Santé
http://www.ecosante.fr/index2.php?base=DEPA&langh=FRA&langs=FRA&sessionid
Institut de recherche et de documentation en économie de la santé http://www.irdes.fr/

Other infrastructures listed in PPP WP10_T4_V3_R1_7.8.09 final.doc

Lists of SSH data producers in each country studied (July 2017)

France

Web-site	Main Subject(s)	Actor	Org_Type	Role	Name
http://www.archivesnationales.culture.gouv.fr/	Labour and Employment, Social Policy and Systems	parallel	LC	DA	Archives nationales (national archives)
http://www.msh-reseau.fr/spip.php?article34	Society and Culture, Social Stratification and Groupings-Inequalities	parallel	Other data org	DA	Archives de la recherche en sciences humaines et socials (ARSHS)
http://www.credoc.fr/	Society and Culture, Social Policy and Systems	parallel	Other data org	PD	Centre de recherche pour l'étude et l'observation des conditions de vie (CREDOC)
http://www.fnors.org/index.html	Health	parallel	Other data org	PD	La Fédération nationale des observatoires régionaux de la santé (FNORS)
http://www.ids.fr	Social Stratification and Groupings-	parallel	Other data org	PD	L'Institut du Développement Social (IDS)

	Inequalities, Health				
http://www.ifop.com/europe	Health, Politics	parallel	Other data org	PD	l'Institut Français d'Opinion Publique (IFOP)
http://www.ign.fr/	Natural Environment	parallel	Other data org	PD-PV	Institut Géographique National (IGN)
http://inpes.santepubliquefrance.fr/	Health	LC contractors	Other data org	PD-PV	Institut national de prévention et d'éducation pour la santé (INPES) (Santé Publique France since 2016)
http://www.inra.fr/	Housing and Land Use Planning, Natural Environment	parallel	Other data org	PD	Institut national de la recherche agronomique (INRA)
http://www.inrets.fr/	Transport, Travel and Mobility	parallel	Other data org	PD	Institut national de recherche sur les transports et leur sécurité (INRETS)
http://www.inserm.fr/fr/	Health, Science and Technology	parallel	LC	PD	l'Institut national de la santé et de la recherche médicale (INSERM)
http://www.ipsos.fr/	Social Stratification and Groupings-Inequalities, Health	concurrent	Other data org	PD: DA	IPSOS
http://www.irdes.fr/	Trade, industry and Markets, Politics	parallel	Other data org	PD	Institut de recherche et de documentation en économie de la santé (IRDES)
http://www.iresp.net/	Health	parallel	Other data org	PD-PV	Institut de recherche en Santé Publique (IRESP)
http://www.mediametrie.fr/	Health	parallel	Other data org	PD	MEDIAMETRIE
http://www.inhes.interieur.gouv.fr/Observatoire-national-de-la-delinquance-6.html	Law, Crime and Legal Systems	parallel	Other data org	PD	Observatoire de la Délinquance
http://www.tns-sofres.com/	Law, Crime and Legal Systems	concurrent	Other data org	PD	TNS SOFRES
http://www.banque-france.fr	Trade, industry and Markets, Politics	parallel	Other data org	PD	Banque de France

Banque de données en santé publique (BDSP)

http://www.bdsp.ehesp.fr/

The device Public Health Database (BDSP) consists of libraries, documentation centres, producers and disseminators of information, the public health field specialists. These resource agencies have chosen to partner to develop, supply and distribute information services in the field of public health. To date, forty members participating in this collective knowledge capitalization

1. Agence régionale de santé Pays de la Loire http://www.ars.paysdelaloire.sante.fr

2. ANFH – Association Nationale pour la Formation Permanente du Personnel Hospitalier http://www.anfh.fr
3. ARAMIS – Réseau des producteurs de données en santé publique
    a. CREAI-ORS Languedoc-Roussillon
    b. CRIPS Ile de France
    c. INPES
    d. IRDES
    e. Observatoire Régional de la Santé d'Auvergne
    f. Observatoire Régional de la Santé Midi Pyrénées
    g. Observatoire Régional de la Santé Rhône-Alpes
    h. SAPHIR (Swiss Automated Public Health Information Resources)
4. ARSI – Association de Recherche en Soins Infirmiers
5. Ascodocpsy – Réseau documentaire en santé mentale
6. ASPBD – Société Française des Acteurs de la Santé Publique Bucco-Dentaire
7. ASPHER – Association des Ecoles de Santé Publique de la Région Européenne
8. Assistance Publique-Hôpitaux de Paris (AP-HP) – Réseau documentaire
    a. APHP IFSI Bichat – Hôpital Bichat – Claude Bernard – Institut de Formation en Soins Infirmiers – Centre de documentation
    b. APHP IFSI Mondor – Hôpital Henri Mondor – Institut de Formation en Soins Infirmiers – Centre de documentation
    c. APHP IFSI Pitié-Salpêtrière – Hôpital Pitié-Salpêtrière – Institut de Formation en Soins Infirmiers
    d. APHPDOC – Centre de Documentation de l'Assistance Publique – Hôpitaux de Paris
9. BIU Santé – Bibliothèque interuniversitaire de santé
10. CCLIN NosoBase – Réseau des Centres de Coordination de Lutte contre les Infections Nosocomiales
    a. CCLIN Ouest – Centre de Coordination de la Lutte contre les Infections Nosocomiales Ouest
    b. CCLIN Paris-Nord – Centre de Coordination de la Lutte contre les Infections Nosocomiales Paris-Nord
    c. CCLIN Sud-Est – Centre de Coordination de la Lutte contre les Infections Nosocomiales Sud-Est
    d. CCLIN Sud-Ouest – Centre de Coordination de la Lutte contre les Infections Nosocomiales Sud-Ouest
11. CERFEP (CEntre de Ressources et de Formation à l'Education du Patient) – CARSAT Nord-Picardie
12. CNSP-FV – Centre national des soins palliatifs et de la fin de vie
13. CREAI-ORS Languedoc-Roussillon
14. CRIPS – Réseau national des Centres Régionaux d'Information et de Prévention du Sida
    a. CRIPS AQUI – CRAES-CRIPS Aquitaine
    b. CRIPS AUV – Centre Régional d'Information et de Prévention du Sida Auvergne (APS-CRIPS Auvergne)
    c. CRIPS IDF – Centre Régional d'Information et de Prévention du Sida Ile-de-France

   d. CRIPS MARS – Centre Régional d'Information et de Prévention du Sida Provence-Alpes-Côte-d'Azur, antenne de Marseille
   e. CRIPS NICE – Centre Régional d'Information et de Prévention du Sida Provence-Alpes-Côte-d'Azur, antenne de Nice
   f. CRIPS RA – Centre Régional d'Information et de Prévention du Sida Rhône-Alpes
15. EHESP – Ecole des hautes études en santé publique
16. ENSP – Ecole nationale de santé publique – Rabat – Maroc
17. Equipe technique de la BDSP
18. ESP – Ecole de Santé Publique
19. ESPRIT – Espace de Santé Publique Régional InTeractif
20. FHF – Fédération hospitalière de France
21. HAS – Haute Autorité de santé
22. HCSP – Haut Conseil de la Santé Publique
23. INAVEM – Institut national d'aide aux victimes et de médiation
24. Inist-CNRS – Institut de l'Information Scientifique et Technique
25. IRDES – Institut de Recherche et Documentation en Economie de la Santé
26. Ministère des affaires sociales et de la santé
27. MSSH-EHESP – Maison des sciences sociales du handicap
28. Observatoire Régional de Santé d'Ile-de-France
29. OFDT – Observatoire français des drogues et des toxicomanies
30. ORS Auvergne – Observatoire Régional de la Santé d'Auvergne
31. ORSMIP – Observatoire Régional de la Santé de Midi-Pyrénées
32. ORSPACA – Observatoire Régional de la Santé de Provence-Alpes-Côte d'Azur
33. ORSPEC – Observatoire Régional de la Santé du Poitou-Charentes
34. ORSRA – Observatoire Régional de la Santé de Rhône-Alpes
35. Santé Publique France
36. SAPHIR – Swiss Automated Public Health Information Resources
37. SFSP – Société Française de Santé Publique
38. Université Paris VII Denis Diderot, UFR Lariboisière Saint Louis, Département de Santé Publique

List of approved hubs (corporate companies, universities, hospitals, etc.) stocking health personal data, having received an agreement from the French government (updated - July 10, 2017)
http://esante.gouv.fr/services/referentiels/securite/hebergeurs-agrees

Website	*Host*
http://www.a2com.fr/cloud-computing	A2COM
http://www.2csi.info/	2CSI
http://www.aatlantide.com/	AATLANTIDE
http://www.abscisse.fr/	Abscisse Informatique
http://www.adista.fr/	Adista
https://www.almerys.com/health/accueil/	Almérys SAS

http://www.arrowecs.fr/	Arrow ECS
http://www.arrowecs.fr/services_1/asplenium/aspleniu m.cfm	Asplénium Hosting Services
http://fr.ap-hm.fr/ap-hm	Assistance Publique des Hôpitaux de Marseille (AP-HM)
www.aphp.fr	Assistance Publique des Hôpitaux de Paris (AP-HP)
http://www.ate.info/	Avenir Télématique (ATE)
http://www.aznetwork.eu/	AZ NetWork
http://www.biotronik.com	Biotronik France
http://www.bull.fr/secteurs/offre-sante.html	Bull
http://www.globalservices.bt.com/fr	BT Services
http://www.carestreamhealth.fr/publicIndex.aspx?Lang Type=1036	Carestream
http://www.cegedim-activ.com	Cegedim Activ'
http://www.cegedim.fr/solutions/Pages/default.aspx	Cegedim SA
www.cegedim.fr/cloudservices	Cegedim SA
http://www.cegialfa.fr/	Cegi Alfa
http://www.cerner.fr/	CERNER
http://www.cev-solutions.com/solutions/cev-sante.html	CEV Group
http://www.cheops.fr/	Cheops Technology
http://www.choregie.fr/	Chorégie
http://www.chu-nantes.fr/	CHU de Nantes
http://www.chu-nice.fr/	CHU de Nice
http://www.chru-strasbourg.fr	CHU de Strasbourg
http://www.chu-nancy.fr	CHRU de Nancy
http://www.chi-eureseine.fr	CHI Eure-Seine
http://www.cimut.fr	CIMUT
http://www.synaaps.com	Ciril GROUP – SynAApS
http://www.cis-valley.fr/	CIS Valley
https://www.claranet.fr/infogerance/hds-esant%C3%A9	CLARANET (anciennement GRITA)
http://www.coaxis-asp.fr/	Coaxis ASP
www.cgm.com/fr	CompuGroup Medical France/Réseau Santé Social
http://www.cgm.com/fr	CompuGroup Medical France/Réseau Santé Social
http://www.dci.fr	Data Concept Informatique (DCI)
http://www.diademys.com/	Diadémys
http://www.docapost.com/	DOCAPOST BPO
www.econocom.com	ECONOCOM-OSIATIS France
http://www.ecritel.fr	ECRITEL
http://www.eig.fr	EIG
http://www.epiconcept.fr/	EpiConcept
http://www.emosist.fr/portail	GCS EMOSIST-FC
https://www.sante-martinique.fr/portail/	GCS SIS de Martinique
http://www.sante-lorraine.fr/	GCS Télésanté Lorraine

https://www.tesis.re	GCS TESIS
www.e-sis5962.fr	GIP e-SIS 59/62
http://www.mipih.fr/	GIP MiPih
http://www.sib.fr/	GIP SIB
http://www.mes-sauvegardes-de-sante.com/	GMI Expert
http://www.bostonscientific.com/en-EU/home.html	Guidant Europe
http://www.h2ad.net/	H2AD
http://www.chu-lyon.fr/	HCL
http://www.ibo.fr/	I.B.O.
http://www.ibm.com/fr/fr/	IBM France
http://www.ids-assistance.com/	IDS
http://www.i-invest.net/	I-Invest
http://www.group-ict.com/	International Cross Talk
http://www.interxion.fr/	Interxion
http://www.locarchives.fr/	Locarchives
http://www.lomaco.fr	Lomaco
http://www.navaho.fr/	Navaho
http://www.netplus.fr/	NetPlus
http://www.numergy.com/	NUMERGY
http://www.orange-business.com/	ORANGE BUSINESS SERVICES
https://www.ovh.com/fr/	OVH
http://www.pacesetter.com/	Pacesetter
http://www.pharmagest.com/	Pharmagest
http://www.pharmagest.com/	Pharmagest
http://www.coreye.fr/fr/silver-sante	Pictime/Coreye
http://www.probtp.com/	PRO BTP
http://www.proginov.com/	Proginov
http://www.prosodie.fr/	Prosodie Capgemini
http://www.runiso.com/	Runiso
http://www.santeos.com/	SANTEOS
http://www.sfrbusinessteam.fr/sante/	SFR
http://www.sigems.fr/	SIGEMS DATA CENTER
http://www.sigma.fr/	Sigma Informatique
http://www.softwaymedical.fr/	Softway Medical Services
http://www.softwaymedical.fr/	Softway Medical Radiologie
http://www.solware.fr/life	Solware Life
http://www.sorin.com/	Sorin CRM
http://www.silpc.fr/	Syndicat Interhospitalier du Limousin
http://www.thalesgroup.com/cic	Thales
http://www.telecitygroup.fr/	TelecityGroup Groupe Equinix
www.tessidocumentsservices.fr	TESSI GED
http://www.fr.zayo.com/	ZAYO France

B. HISTORICAL DATA – STATE OF PLAY/OTHER ACTORS

- **Slovenia**. Slovene historiography portal –Sistory.

- **Czech Republic**. a) Bavarian-Czech network of digital historical sources, b) National Archive.

- **Denmark**. a) The Danish Emigration Archives, b) the Danish State Archives Filming Centre.

- **Switzerland.**a) Infoclio - the professional portal of the historical sciences in Switzerland. b) Historical Statistics of Switzerland Online.

- **Lithuania**. a) Lithuanian State Historical Archives.

- **France**. a) National Institute of Statistics and Economic Studies (Insee), b) The Diplomatic Archive Centre of the Ministry of Foreign and European Affairs. c) The Defence Historical Service (*Service historique de la défense - SHD*) is the archives centre of Ministry of Defence and its' armed forces. The SHD consists of the "Centre historique des archives" at Vincennes, the "Centre des archives de l'armement et du personnel" at Châtellerault, the Archives de la Fondation Maison des sciences de l'homme and a number of smaller repositories.

- **Hungary.** a) The Historical Archives of the Hungarian State Security**,** b) The Hungarian National Archives, c) Ecclesiastical Archives.

- **Sweden**. National Archives.

- **Finland**. a) University of Turku, Department of History, b) ÅboAkademi,c) Finish American Historical Archive.

- **Germany.** a) Deutsches Historisches Museum (Historical Museum), b) Bundesarchiv (German Federal Archives), c) Württembergische Landesbibliothek Stuttgart (Library of Contemporary History), d) Hamburger Institut für Sozialforschung, Institut für Zeitgeschichte.

- **United Kingdom**. The National Archives.

- **Netherlands**. a) National Archives of the Netherlands, b) The Stichting Mondelinge Geschiedenis Indonesië (SMGI), Koninklijk Instituut voor Taal and the Veteranen Instituut, which have deposited more than 1,000 oral history datasets in DANS.

- **Norway.** a) NTNU Historical Archives, b) The National Archives of Norway, c) The Norwegian Historical Data Centre (NHDC).

- **Greece**. a) Historical Archives of Museum Benaki, b) General State Archives, c) Archives of the Hellenic Army, d) Historical Archive of the National Bank, e) Historical Archive of the Communist Party, f) Historical Archive of the Bank of Greece, g) Historical Archive of the Foundation K.G. Karamanlis, h) Institute of Historical Research, National Hellenic Research Foundation, i) Archives of Modern Social History, k) Centre for Asia Minor Studies.

APPENDIX II.

Interviews protocols

Interview guide for big data

A. Definition of big data, namely the areas of research interests of the domain

Indicative Questions:

Q.1. How can we approach the field of big data? Please, describe the basic categories.

B. Scientific networks (inside-outside the country).

Indicative Questions:

Q.1. Are you member of national or international scientific networks regarding the production, preservation and/or dissemination of big data?

C. Dissemination of big data.

Indicative Questions:

Q.1. How you disseminate big data resulting from your own research projects?

D. Main databases for big data. Advantages and insufficiencies.

Indicative Questions:

Q.1. Can you mention some of the main databases providing big data from which you retrieve data? Which actors /agencies hold relevant databases?

Q.2. Do you know if the most important databases implement some common classification standards in order to facilitate researchers' work?

Q.3.Can you name some health databases which you perceive as well structured, regarding researchers' ability to easily access and retrieve data?

Q.4. based on your experience are there any constrains /obstacles that should be overcome in order to facilitate researchers' access (i.e metadata etc)?

E. Current and future researchers' needs for big data.

Q.1. Do you believe that researchers can easily locate the majority of databases providing big data? If not, please cite some reasons.

Q.2. Which sectors of big data currently present the greatest interest for researchers and/or national and international organisations? Why? Which ones do you expect to present the greatest interest in the near future?

Interview Guide for Health data

A. Definition of health data namely the areas of research interests of the domain eg. diseases, bioresearch, drags use/abuse, epidemiological studies etc.

Indicative Questions:

Q.1. How can we approach the field of "health data"? Please, describe the basic categories.

B. Scientific networks (inside-outside the country).

Indicative Questions:

Q.1. Are you member of national or international scientific networks regarding the production, preservation and/or dissemination of health data?

C. Dissemination of health data.

Indicative Questions:

Q.1. How you disseminate health data resulting from your own research projects?

D. Main databases for health data. Advantages and insufficiencies.

Indicative Questions:

Q.1. Can you mention some of the main health databases from which you retrieve data? Which agencies hold relevant health databases?

Q.2. Do you know if the most important databases implement some common classification standards in order to facilitate researchers' work?

Q.3.Can you name some health databases which you perceive as well structured, regarding researchers' ability to easily access and retrieve data?

Q.4. based on your experience are there any constrains /obstacles that should be overcome in order to facilitate researchers' access (i.e metadata etc)?

E. Current and future researchers' needs for health data.

Q.1. Do you believe that researchers can easily locate the majority of health databases? If not, please cite some reasons

Q.2. Which sectors of health data currently present the greatest interest for researchers? Which ones do you expect to present the greatest interest in the near future?

Interview Guide for Historical Data

A. Definition of Historical data

Indicative Questions:

Q.1. How can we approach the field of historical data? Please, describe the basic categories of historical data.

Q.2. Which are the boundaries between the humanities and the social sciences?

B. Scientific networks (inside-outside the country).

Indicative Questions:

Q.1. Are you a member of national or international scientific networks regarding the production, preservation and/or dissemination of historical data?

C. Dissemination of historical data.

Indicative Questions:

Q.1. How you disseminate historical data resulting from your own research projects?

D. Main databases for historical data. Advantages and insufficiencies.

Indicative Questions:

Q.1. Can you mention some of the main historical databases from which you retrieve data? Which actors /agencies hold relevant databases?

Q.2. Do you know if the most important databases implement some common classification standards in order to facilitate researchers' work?

Q.3.Can you name some historical databases which you perceive as well structured, regarding researchers' ability to easily access and retrieve data?

Q.4. based on your experience are there any constrains /obstacles that should be overcome in order to facilitate researchers' access (i.e metadata etc)?

E. Current and future researchers' needs for historical data.

Q.1. Do you believe that researchers can easily locate the majority of historical databases? If not, please cite some reasons

Q.2. Which sectors of historical data currently present the greatest interest for researchers? Which ones do you expect to present the greatest interest in the near future?